DOT/FAA/AR-96/65

Office of Aviation Research
Washington, D.C. 20591

Visual Inspection Research Project Report on Benchmark Inspections

October 1996

Final Report

U.S. Department of Transportation
Federal Aviation Administration

NOTICE

This document is disseminated under the sponsorship of the U.S. Department of Transportation in the interest of information exchange. The United States Government assumes no liability for the contents or use thereof. The United States Government does not endorse products or manufacturers. Trade or manufacturer's names appear herein solely because they are considered essential to the objective of this report.

Technical Report Documentation Page

1. Report No.	2. Government Accession No.	3. Recipient's Catalog No.
DOT/FAA/AR-96/65		

4. Title and Subtitle	5. Report Date
VISUAL INSPECTION RESEARCH PROJECT REPORT ON BENCHMARK INSPECTIONS	October 1996
	6. Performing Organization Code

7. Author(s)	8. Performing Organization Report No.
Floyd W. Spencer	DOT/FAA/AR-96/65

9. Performing Organization Name and Address	10. Work Unit No. (TRAIS)
FAA Aging Aircraft NDI Validation Center Sandia National Laboratories P.O. Box 5800 Albuquerque, NM 87185-0829	11. Contract or Grant No.

12. Sponsoring Agency Name and Address	13. Type of Report and Period Covered
U.S. Department of Transportation Federal Aviation Administration Office of Aviation Research Washington, DC 20591	Final Report
	14. Sponsoring Agency Code AAR-433

15. Supplementary Notes

FAA William J. Hughes Technical Center COTR: Christopher Smith

16. Abstract

Recognizing the importance of visual inspection in the maintenance of the civil air fleet, the FAA tasked the Aging Aircraft Nondestructive Inspection Validation Center (AANC) at Sandia National Labs in Albuquerque, NM, to establish a visual inspection reliability program. This report presents the results of the first phase of that program, a benchmark visual inspection reliability experiment. The benchmark experiment had 12 airline inspectors perform specific inspection tasks on AANC's Boeing 737 in order to estimate overall performance characteristics of a typical set of inspectors on a typical set visual inspection tasks. The report also includes a separate but related probability of detection study on small but visible cracks at rivet locations on fabricated fuselage skin splices. Conclusions are drawn with respect to quantification of inspection reliability, search and decision aspects of visual inspection, use of job cards during inspection, and inspector specific factors affecting visual inspection performance.

17. Key Words		18. Distribution Statement
Visual inspection Nondestructive inspection Human factors	Inspection reliability Probability of detection Job card	This document is available to the public through the National Technical Information Service (NTIS), Springfield, Virginia 22161.

19. Security Classif. (of this report)	20. Security Classif. (of this page)	21. No. of Pages	22. Price
Unclassified	Unclassified	59	

PREFACE

The Aging Aircraft Nondestructive Inspection Validation Center (AANC) was established at Sandia National Laboratories by the Federal Aviation Administration (FAA) in August 1991. This center is dedicated to the study and validation of nondestructive inspection techniques and technologies and is housed in a hangar at the Albuquerque International Airport. The AANC possess a number of aircraft, aircraft parts or components, and specially constructed test pieces for this purpose.

The FAA Interagency Agreement, which established the AANC, provided the following summary tasking statement: "The task Assignments will call for Sandia to support technology transfer, technology assessment, technology validation, data correlation, and automation adaptation as ongoing processes." In short, Sandia National Laboratories has been asked to pursue research to improve nondestructive inspection (NDI) for the aging aircraft program. Recognizing the importance of visual inspection in the maintenance of the civil air fleet, the AANC established a Visual Inspection Reliability Program. This report presents the results of the Benchmark phase of that program. The Benchmark consisted of obtaining inspection results from twelve experienced inspectors on AANC's Boeing 737. All the inspectors used the same job cards and inspected the same areas of the test bed at the AANC.

Various organizations have helped the AANC in planning and executing the Benchmark phase reported here. The principle investigator at the AANC was Floyd Spencer. Assisting in the program and in the writing of this document was Donald Schurman, formerly of Science Applications International Corporation; Colin Drury, State University of New York at Buffalo; Ron Smith and Geoff Worrall, AEA Technology.

We especially thank the Inspection Network members of the Air Transportation of America (ATA) for their assistance and guidance during the planning phases of this Benchmark experiment. Special thanks to Steve Erickson, formerly of the ATA; Bob Scoble, United Airlines; Roy Weatherbee, USAir; John Spiciarich, TWA; Mike Guiterrez, Federal Express; and Hugh Irving, Delta; for attending meetings and providing direct input. Ward Rummel of Martin Marietta, also contributed to the planning.

Any exercise of this nature would have no hope of success without the willing participation of the airline inspectors who not only spent two days performing inspections and answering questions but also allowed themselves to be videotaped. A very special thanks to the twelve inspectors that participated in the Benchmark inspections. They were employed at Continental Airlines, Delta Airlines, United Airlines, and USAir, and we thank their management for their assistance in obtaining the services of the inspectors.

Thanks are also extended to John Goglia and the International Association of Machinists and Aerospace Workers (IAM) District 141 Flight Safety Committee. Twenty-four members of the committee graciously volunteered an hour of their time to participate in a flashlight/crack detection experiment while touring the AANC.

Thanks are also extended to Pat Walter, Texas Christian University, who, at the inception of the program, was the manager of the AANC and was instrumental in establishing the program. Chris Smith, William J. Hughes Technical Center, was the technical monitor for the project and assisted throughout the project.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

EXECUTIVE SUMMARY

Visual inspection is the first line of defense for safety-related failures on aircraft and provides the least expensive and quickest method of assessing the condition of an aircraft and its parts. As such, its reliability should be high and well-characterized. This report describes the Benchmark Experiment of the Visual Inspection Research Program performed at the FAA's Aging Aircraft Nondestructive Inspection Validation Center (AANC) at Sandia Laboratories. The purpose of the experiment was to provide a benchmark measure of capability for visual inspections performed under conditions that are realistically similar to those usually found in major airline maintenance facilities.

Most of the research related to visual inspection has been in the area of visual search conducted for industrial products and medical images. This research is reviewed here, but the intent was to target aircraft specific visual inspections over a variety of tasks. The tasks chosen represent different accessibility levels, as well as different visual complexity levels.

The research described here is the first part of a coordinated effort to examine the broad range of visual inspection requirements. It is neither completely a laboratory study nor completely field observations. Instead, it provides a link between field and laboratory by using visual inspection tasks on a real airplane combined with other more controlled tasks. The AANC has a Boeing 737 aircraft as a test bed. In addition, the AANC has a sample library of well-characterized flaws in aircraft components or simulated components that allow cross-linking of the aircraft results with well understood flaw characteristics. Both of these resources were used for this research.

Twelve inspectors from four different airlines served as the subjects of the experiment. Each subject spent two days at the AANC performing 10 different inspection tasks. Data collection consisted both of notes taken by monitors and videotaping of the inspection tasks. Performance results are summarized and correlated with background variables gathered on each inspector.

Substantial inspector-to-inspector variation in performance was observed. This observation has direct bearing on determining sample sizes necessary to study the impact of visual inspection factors or the effectiveness of specific interventions. On a specific task of looking for cracks from beneath rivet heads the 90 percent detection rate crack length for 11 inspectors ranged from 0.16 to 0.36 inches, with the 90 percent detection rate for the twelfth inspector being 0.91 inch. Similar variations were observed in other inspection tasks.

Performance levels were task specific. Thus, an inspector's good performance (relative to other inspectors) on one task does not necessarily indicate a relatively good performance on other tasks. The search component, as opposed to the decision component, of visual inspection was the larger factor in determining performance levels for most of the inspectors.

Major factors associated with performance in this research were the use of job cards, thoroughness as reflected by total time, peripheral visual acuity, and general aviation experience. Specifically, some of the inspectors used job cards only to define the inspection task area. Others used the information contained within the job card to direct their attention to likely flaw locations. The inspectors with lower peripheral visual acuity scores showed a decline in performance on certain tasks. Better performances were observed among inspectors with more aviation experience as well as those that were more deliberate in their inspections as reflected by the time taken on all tasks.

1. INTRODUCTION.

This report describes a Benchmark visual inspection experiment and a flashlight lens evaluation experiment. Both of these activities were carried out as part of the Visual Inspection Research Program being conducted at FAA's Aging Aircraft Nondestructive Inspection Validation Center (AANC) at Sandia Laboratories.

The Benchmark experiment provided a measure of capability for visual inspections performed under conditions that are realistically similar to those found in major airline maintenance facilities. The characterization of the visual inspection process from this benchmark serves as a measure to compare performance under other conditions. The variations observed in the Benchmark experiment also enable an assessment of the amount of testing necessary to measure inspection performance effects due to environments, conditions, and instructions.

The flashlight lens experiment was a follow-on to a flashlight lens development program described by Shagam [1]. The subjects for this experiment were drawn from aircraft inspectors and mechanics. The test specimens used in the flashlight lens experiment were also used in the Benchmark experiment. The intent of the flashlight lens experiment was to study the len's effect on subject's performance. However, due to the similarity of subject populations and the commonality of test specimens, this experiment complements the larger Benchmark experiment and results are included in this report.

1.1 BACKGROUND.

Over 80 percent of inspections on large transport category aircraft are visual inspections [2]. Small transport and general aviation aircraft rely on visual inspection techniques even more heavily than do large transport aircraft. Visual inspection, then, is the first line of defense for safety-related failures on aircraft and provides the least expensive and quickest method of assessing the condition of an aircraft and its parts [3]. Therefore, accurate and proficient visual inspection is crucial to the continued safe operation of the air fleet.

1.2 DEFINITION OF VISUAL INSPECTION.

Visual inspection has been defined in one FAA publication [3] as, "... the process of using the eye, alone or in conjunction with various aids, as the sensing mechanism from which judgments may be made about the condition of a unit to be inspected." This definition is good—as far as it goes, however, experience in visual inspection and discussion with experienced visual inspectors reveals that this definition falls short. Not only does visual inspection involve the use of the "eye, alone or with various aids" but also involves, shaking, listening, feeling, and sometimes even smelling the aircraft and its components.

Thus, the definition of visual inspection must include the use of the other senses as well. Visual inspection (and other types of inspection, as well) consists of at least two major processes [4]. The first is a search process that, in visual inspection, uses most of the senses of the human body. The second process is a process of combining relevant knowledge, sensory input, and pertinent logical processes to provide an identification that some anomaly or pattern represents a flaw and

logical processes to provide an identification that some anomaly or pattern represents a flaw and a decision that this flaw is of the nature to pose a risk to the continued successful operation of the aircraft or aircraft part.

For the Visual Inspection Research Program, we expand the definition of "Visual Inspection" to include other sensory and cognitive processes that are used by inspectors. We feel that neglect of these other systems provides an artificially narrow picture of the rich range of behaviors involved in visual inspection. Thus, the Visual Inspection Research Program uses the following definition of Visual Inspection:

> Visual inspection is the process of examination and evaluation of systems and components by use of human sensory systems aided only by such mechanical enhancements to sensory input as magnifiers, dental picks, stethoscopes, and the like. The inspection process may be done using such behaviors as looking, listening, feeling, smelling, shaking, and twisting. It includes a cognitive component wherein observations are correlated with knowledge of structure and with descriptions and diagrams from service literature.

1.3 VISUAL INSPECTION RESEARCH.

This section provides an abbreviated summary of previous research related to visual search in general and to aircraft inspection in particular.

1.3.1 Visual Search—General.

The basis for scientific research applied to visual inspection tasks has been primarily the study of visual search. This research has attempted to extrapolate the findings of laboratory-based studies to visual inspection tasks that are found in the quality assessment systems of most manufacturing industries.

As a summary of this research, visual search is considered to be a series of short fixations by the eye during which information is gathered. These fixations are interspersed by rapid eye movements in which the area of fixation is moved to another part of the object being viewed. The area that surrounds the fixation point, and from which the eye collects information, is called the visual lobe. (This visual lobe is elliptical but is treated as a circle for convenience.) The boundary is defined by the angle from the center of fixation which allows a 50 percent detection rate. The size of the target being searched for, the level of contrast, and the luminance level of the background all directly affect the detection rate of a target [5].

If targets are changed to provide a larger surface area (while maintaining aspect ratios), the probability of detection is improved. Increased edge sharpness also has the same effect. If a larger region is to be searched, in a fixed time, then the probability of detection is reduced. In addition, if the prior expectation of a target occurring is increased, so does the probability of detection [6].

The speed at which visual search (target detection and localization) can take place is affected by four factors [7]:

- *Number of elements to be searched*, i.e., as the number of items goes up so does search time, relatively independent of element spacing. Wide dispersal of elements increases scanning time. However, when items are closely packed the high density of nontarget elements also has a retarding influence on search. Thus, scanning and visual clutter trade off as target dispersion is varied.

- *Search rate increases as the total amount of information in the display increases.* Information is increased by increasing the number of items to be searched, the number of variable stimulus dimensions per item, or the number of possible targets. However, the increase in search rate (items per unit time) with more items does not compensate for the increased time needed to search more items. As a result total search time is increased by including more items or more relevant dimensions per item.

- *Searching for one of several targets is slower than searching for one.* Some laboratory based results have not shown this effect if extensive training is given. This would suggest a well trained inspector would not be slowed by greater numbers of targets to search for, but this is not conclusive [8, 9].

- *Number of different stimulus dimensions that can be used to define a target does not affect speed if they are redundant.* For example, color is a more salient dimension than is shape, and thus, searching for blue squares and blue circles can be done as quickly as just searching for blue squares, as long as nontargets are not blue. However, color will interfere with perception of other dimensions, such as shape, if it varies independently of them.

When time is limited for visual search tasks, it has been shown that fixating patterns adopted by subjects will be altered [10]. Rather than moving from target to target, subjects will attempt to get a larger visual field and not move the eyes. It has also been demonstrated that experienced inspectors use a different visual search strategy than do inexperienced inspectors [8], although there is also some evidence to suggest that trained inspectors do not perform more effectively than untrained subjects on visual search tasks [9]. Limiting the time (increasing the speed) of industrial visual search tasks also reduces the probability of accepting good items while raising the probability of rejecting bad items [11].

Conscious direction of attention to specific areas is possible based upon the inspector's previous experience. However, beyond this conscious direction, within the visual lobe, an automatic mechanism of selective attention seems to apply with salient target characteristics dominating attention. In addition to these visual factors, there is evidence that the psychological profile of subjects will also have an affect on visual search performance [12].

A list of the factors that have been tested and concluded to affect inspection tasks has been compiled [13]. This list does not identify at what stage within the inspection task these factors

are said to affect the inspection performance however. The identified factors can be broadly split into four areas: subject factors, task factors, organizational factors, and physical and environmental factors.

There have been at least two separate descriptions of the components of visual inspection. Splitz and Drury [14] model visual inspection in two separate stages, search and decision making. The model proposes that these two stages are separate and additive and goes on to provide experimental evidence to support this view.

Megaw's model suggests that there are four separate stages [15]:

- Search: Scanning item via head and eye and hand movements (moving the object).
- Detect: Identify that the item is different from its ideal state.
- Judgment: Decide whether the difference constitutes a fault according to the standards to which the task is being performed.
- Output decision: Decide whether to accept or reject and take the appropriate action.

It can be argued that this second view is not different from the first, but rather, that it separates the two stages of the first model into two subcomponents (that is, "search" into "search" and "detection" and "decision" into "judgment" and "output decision").

Studies of eye movement during industrial and medical x-ray inspection [15] have shown:

- Inspection time differences reflect the number of fixations needed to search for and find a fault rather than differences in the time of each fixation.
- Fixation times are short in tasks without clear fixation points (e.g., inspection of sheet steel) and that more experienced inspectors use shorter fixation times. Fixation time with respect to task complexity has not been well-explored.
- In objects which were manipulated, scan paths were fixed and errors occurred as a result of sticking to these scan paths.
- Peripheral vision is used for scanning moving objects which subtend a large visual angle.

This body of work provides a foundation on which to base the more applied studies carried out in this program. As stated above, the previously described work has centered around theories of visual search. The summarized work largely reflects the nature of the visual inspection tasks carried out by inspectors in the manufacturing industry, which historically has centered around the visual observation of a limited range of items, for example products or x-ray images. Typical

aircraft inspection tasks, however, involve a significantly increased use of other senses and level of manipulation. Consequently the need for a more applied study has arisen for aircraft inspection tasks.

1.3.2 Aircraft Inspection.

There has been some research investigating aircraft inspection tasks specifically as opposed to visual inspection in general. There have also been a small number of studies which have investigated the sensitivity of inspectors to the types of tactile cues which are found in aircraft inspection [16, 17, 18]. Three distinct types of visual inspection tasks were highlighted by one source [19]:

- *Detection:* For example, identification of a warning light or breaks in a seal. In this type of task the inspector only needs to see an object against a background.

- *Recognition:* The inspector needs to detect a stimulus and then discriminate against other possibilities. This type of task has a cognitive component as it involves a comparison to decide if what has been observed constitutes a flaw. This may necessitate better sensory conditions to allow an appropriate level of perception.

- *Interpretation:* Further actions are necessary following the recognition of stimuli. Knowledge of component function and system integration play a role. This task involves much greater cognitive behavior than simply having a sensory system capable of making a discrimination necessary for judgment. Visual inspection in aircraft maintenance falls into this type, as exemplified by the evaluation of "smoking" rivets where an inspector will consider numerous factors, including color and location, to determine if a problem exists.

An overview of the research being undertaken in the area of visual inspection in aircraft maintenance can be found in reference 20. Areas noted specific to aircraft inspection performance include training, information systems design, and international differences [20]. These are briefly discussed in the following paragraphs.

- *Training:* The training of inspectors is a major determinant of inspector performance. At present there is an emphasis on either the classroom for imparting knowledge and the actual job for imparting skills. Little formal training in visual search is given to aircraft inspectors.

- *Information systems design:* The information presented to the inspectors can occur in several ways. Examples are:

  - Directive information (from training and job card)

  - Feedforward (types and locations of faults expected on this aircraft at this time)

- Feedback (from sample quality control checks of inspected aircraft)

- Error control (error taxonomies are being developed and applied to detailed analysis of the inspection system [21, 22]).

- *International differences*: Comparisons between the United Kingdom and the United States systems of aircraft inspection have also been documented [23]. Inspection/maintenance system and hangar floor operations were compared. Differences were found, with probably more variability between companies than between the two countries. Both countries' work forces were found to be highly motivated.

Organizational factors in airline maintenance were described by Taylor [24]. He concluded that the communication of a responsible maintenance role (clear mission statement) within a larger company is usually missing and that maintenance personnel often lack sufficient technical knowledge and have little opportunity to improve decision making and problem solving capabilities. He also concluded that organizational structures within the airline industry emphasized "functional silos," with individual departments working to their own limited goals. Since that time (1990) there have been a number of organizational developments in aircraft maintenance and inspection, often applying Crew Resource Management (CRM) ideas [25].

Research specific to the detection of cracks in aircraft structure was reported by Endoh et al. [26]. In that study, factors associated with cracks found in routine maintenance and inspection activities for aircraft operated by Japanese airlines were documented over a 3-year period. Factors were studied by generating a normalized cumulative frequency with respect to crack lengths for each level of a given factor and graphically comparing these curves. Endoh's data are compared to data from the current study in section 5.6 following a discussion of results for the Benchmark experiment.

1.4 BENCHMARK RESEARCH.

The current research program was formulated to carry out an applied investigation of aircraft visual inspection using the research summarized in sections 1.3.1 and 1.3.2 to ensure that appropriate variables were controlled and appropriate measures taken. As implied by the expanded definition of visual inspection (section 1.2), understanding of the process of visual inspection requires research into the use of other sensory systems in addition to just the visual system. Also, both the search and the decision-making aspects of visual inspection require examination [4].

The research described here is the first part of a coordinated effort to examine the broad range of visual inspection requirements. It is neither completely a laboratory study nor completely field observations. Instead, it provides a link between laboratory and field by using visual inspections on a real airplane combined with other controlled tasks. The FAA Aging Aircraft Nondestructive Inspection Validation Center (AANC) at Sandia Laboratories has a Boeing 737 aircraft test bed. In addition, the AANC has a sample library of aircraft components or simulated components with

well-characterized flaws that allows cross-linking of aircraft inspection results with well understood flaw characteristics.

This first experiment was planned as a field benchmark. That is, the intent was to provide a benchmark of visual inspection performance under realistic conditions. This study looked at the performance and general characteristics of a sample of visual inspectors. Their performance serves as the control, or benchmark, for comparisons to the manipulated or selected characteristics of inspectors in later studies.

Specifically, the study was done using a representative sample of visual inspectors in aircraft maintenance facilities, who were asked to look for flaws on the Boeing 737 test bed aircraft in selected, specific areas that roughly corresponded to normal inspection task requirements. The inspectors were asked to inspect the aircraft for about one and one-half shifts. For the other half shift, the inspectors were asked to look for flaws on a selected set of samples from the AANC sample library. The implementation and logistics are discussed in section 3.

The Benchmark experiment was designed with the assistance and recommendations of an industry steering committee. The committee was formed to provide input and advice on the Visual Inspection Research Program (VIRP) in general and on the Benchmark experiment specifically.

The first planning meeting for the Visual Inspection Research Program (VIRP) was attended by human factors specialists from several organizations and AANC specialists in research, aircraft inspection, and optics. The AANC specialists also had close familiarity with the Boeing 737 in the AANC hangar. This group discussed and developed the broad outline of the VIRP. In broad outline, the VIRP was to start with an experiment to provide a benchmark of visual inspection performance under realistic conditions. Later experiments could be used to test the effectiveness of various interventions on inspection reliability. These interventions might include such actions as improved lighting systems or devices, improved training, improved job card descriptions, improved working conditions and tools, or inspectors with different backgrounds.

The second meeting was also attended by representatives from the Air Transport Association of America (ATA), who were members of inspection and maintenance organizations for large transport carrier fleets. These representatives ensured that practical aspects steered the planning process as well as providing realistic reports of conditions and problems of visual inspection for their organizations' fleets. The ATA representatives were briefed on the VIRP broad outline and the design of the Benchmark experiment. They were also asked to provide input and assistance in determining such factors of the Benchmark as standardized tool kits, structure of job cards, and the choice of initial inspectors.

1.5 FLASHLIGHT LENS EXPERIMENT.

In another Visual Inspection Program project conducted by AANC, an improved flashlight lens which had a pattern molded into one surface to act as a diffuser was developed [1]. The effect is to enhance the uniformity of illumination across the output beam of the flashlight, eliminating

dark and bright spots. This also necessarily reduced the peak brightness. Prior to the experiment presented here, evaluation of this lens had been through the measurement of its optical characteristics and tests of acceptability to practicing inspectors [1]. While both of these are necessary first steps to ensure that the new lens can be used by the industry, they do not address the question of performance. Does this lens aid (or possibly hinder) detection of defects? There have been a number of evaluations of lighting effectiveness in the literature [27]. Typically, changes in lighting must be quite dramatic to achieve significant gains in inspection performance. Merely increasing overall illumination on the task rarely produces performance improvements, unless the original level of illumination was extremely low [28].

The choice of representative people to perform the inspections is critical to experiment validity. A half-day tour of the AANC facility by a committee of the International Association of Machinists and Aerospace Workers (IAM) was accompanied by a request for "hands on" demonstrations. One of the planned demonstrations during their visit was of the light shaping diffuser flashlight lens. Agreement was obtained from the IAM organizing committee that the hands-on experience of the visiting members would be through their participation in a planned experiment to study performance levels associated with the lens in a specific task.

2. DESIGN OF BENCHMARK EXPERIMENT.

2.1 TEST BED.

There were two test beds used in the VIRP Benchmark Experiment. The first was a Boeing 737 aircraft, mostly intact and stripped for a D-check, had certain avionics and one engine removed. The aircraft was put in service in 1968 and had 46,358 cycles in more than 38,000 hours of flight time. It was retired from active service in 1991 after a decision that it would not be economical to bring the aircraft into compliance with safety requirements and airworthiness directives. As part of the activities preparing the B-737 as a test bed for AANC programs, a baseline inspection, consisting of a limited D-check by contract inspectors, was performed. These inspections were completed prior to the start of the VIRP studies. The result of this baseline experiment was a list of defects and flaws found on the aircraft. This list was used to determine the inspection areas to be used in the current and future VIRP studies.

The second test bed consisted of selected flawed specimens from the AANC Sample Library. The samples in the AANC library that were selected were cracked specimen panels and coupons as described by Spencer and Schurman [29]. The samples were well characterized so that the sizes, orientation, location, and distribution of cracks are known. Inspection performance on these samples can be correlated with crack lengths.

2.2 FLAW TYPES.

Although cracks and corrosion are a major concern in aircraft inspection, they are not the only focus of visual inspection. Out-of-tolerance wear or breakage and fraying are also important defects to be detected for safe operation. Identifying wear and tear problems requires detection of a possibly out-of-tolerance condition followed by measurement of dimensions to determine whether the dimensions are within tolerance or require repair or replacement.

The variety of fault types required to be detected is a major factor influencing performance in visual inspection. Another influencing factor is the difficulty of characterization of many fault types. In contrast, the use of nondestructive inspection equipment (such as eddy current and ultrasonic) is usually defect specific with faults being characterized along one or two dimensions, such as crack length, crack width, or area of delamination.

In order to evaluate performance across a range of visual inspection conditions, various flaws requiring a range of behaviors are needed. That is wear and tear defects that require shaking and feeling of components would be required as well as cracks and corrosion. Also, to grade performance, flaws should range from minimally detectable to the trained eye to clearly detectable by a casual observer.

The baseline inspection of the B-737 documented cracks, corrosion, and other types of flaws. A classification scheme was developed to characterize the baseline findings with respect to flaw type as well as with respect to aircraft structure containing the flaws. Flaws within major categories (missing parts, wear and tear, corrosion, disbonds and delamination, wrong part/bad repair, cracks) were further classified with respect to the visually observed condition. For example, flaws within the "wear and tear" category could be further classified in several categories such as loose, frayed, and scratched. Major categories of structure (e.g., skin, fasteners, straps, paints/sealants) as well as subcategories (e.g., internal or external for skins, bolts, rivets or screws for fasteners) were also associated with the various flaws discovered during the baseline inspection. This information was part of the criteria used for choosing tasks for the Benchmark experiment.

2.3 TASK TYPES.

A range of defects in several different locations could involve different levels of physical and visual accessibility. Tasks were selected to cover a range of accessibility (both visual and physical). As far as possible, accessibility was systematically varied. There is not a universally accepted metric for physical or visual accessibility, so we used a post-inspection debrief to get difficulty ratings on accessibility from the inspectors.

A second question that was addressed is the problem of visual complexity, which is only loosely correlated with visual accessibility. Visual complexity refers to the fact that a 1/2-inch-long crack does not have the same detectability when it is located on a lap-splice of the fuselage as when it is located on the inside of a door, surrounded by wire bundles, structural members, and other components. Again, since there is no universally accepted metric for specifying this complexity dimension, the inspectors rated each task for visual difficulty in a post-inspection debriefing.

These task types are not the only factors of importance. Other factors include whether the task is a straightforward visual search, requires shaking and listening, requires feeling for excessive play, etc. That is the flaw type and component type interact with inspection procedures, component location, etc., to determine task type. The Benchmark experiment was planned to sample a range of these task types.

2.4 JOB CARD DESCRIPTIONS.

The consideration of flaw types (section 2.2) as well as task types (section 2.3) and the information available from the baseline inspection led to the selection of the following to constitute the job cards (JC) guiding the inspections of the participating inspectors.

- JC 501—Midsection Floor. Inspection of the midsection fuselage floor beams from body station (BS) 520 to the aft side of BS 727. It included web, chord, stiffeners, seat tracks, upper flanges of floor beams, and over-wing stub beams at BS 559, 578, 597, 616, and 639.

- JC 502—Main Landing Gear Support. Inspection of the main landing gear support fittings for left and right main landing gear support beam and side strut attachments at BS 685, 695, and 706 for cracks and corrosion.

- JC 503—Midsection Crown (Internal). Inspection of the internal midsection crown area stringers and frames from BS 540 to BS 727A from stringer 6L to 6R and tie-clips from stringer 7L to 7R for cracks and corrosion.

- JC—Galley Doors (Internal) Inspection of the galley door frames, hinges, latches, locks, seals, actuating mechanisms, stops and attachments for cracks, corrosion, and general condition (i.e., wear, deterioration). (The galley doors are the two doors on the right side of the aircraft.)

- JC 505—Rear Bilge (External). Inspection of the rear external belly area from BS 727 to BS 907 between stringers 25R and 25L, including lap-splices, for bulges in skin, skin cracks, dished/deformed or missing rivet heads, and corrosion.

- JC 506—Left Forward Upper Lobe. Inspection of the interior of the left fuselage upper lobe from BS 277 to BS 540 and from stringer 17L (floor level) to stringer 4L for corrosion, cracks, and general condition.

- JC 507—Left Forward Cargo Compartment. Inspection of the interior of the left fuselage lower lobe from BS 380 to BS 520 from stringer 18L to the keel beam (centerline) for corrosion, cracks, and general condition.

- JC 508/509—Upper and Lower Rear Bulkhead Y-Ring. Inspection of the aft side of the Y-ring of the fuselage bulkhead at BS 1016 (aft pressure bulkhead) including bulkhead outer ring, Y-frame aft chord, steel strap and fastener locations on all stringers for cracks, corrosion, and accidental damage such as dents, tears, nicks, gouges, and scratches.

- JC 510—Nose Wheel Well Forward Bulkhead. Inspection of the aft and forward side of the nose wheel well forward bulkhead at BS 227.8 for cracks.

- JC 701—Simulated Lap-Splice Panels. Inspection of 38.5 feet of simulated Boeing 737 lap-splice in two types of specimens. The two types consisted of one large (8.5 feet long)

unpainted panel and 18 small panels, each 20 inches long, that were butted against each other and presented as a continuous lap-splice. A description of the test specimens is given in Spencer and Schurman [29].

The job cards were originally numbered 501 through 510 and 701. Job Cards 508 and 509 were inspection of the top and bottom rim (Y-ring and fittings) of the aft side of the pressure dome. They were separated because it was felt that the physical discomfort and visual complexity levels were quite different in the two areas. However, it was found that each job card was so short that it made little sense to do one part of the pressure dome, climb out of the tail cone, and sometime later climb back in to do the other part of the pressure dome. So, the two job cards were combined and were always done together as job card 508/509.

Job Card 502 originally called for inspection of the main landing gear (MLG) support on both sides of the airplane. However, time considerations for the first inspector indicated that this would take too long. Therefore, all the inspectors were asked to inspect only the left side MLG support structure. The job card was modified accordingly for subsequent subjects.

2.5 INSPECTORS.

The inspectors in the Benchmark experiment were all qualified as visual inspectors by their respective airlines (USAir, United Airlines, Delta Airlines, and Continental Airlines) and were working as visual inspectors in their respective facilities. The inspectors from one airline were not always employed at the same facility; seven different maintenance facilities were represented by the 12 inspectors.

2.6 INSPECTION CONDITIONS.

2.6.1 Constant Conditions.

Some conditions that are known or presumed to affect visual inspection performance were standardized or held constant for the Benchmark experiment. These conditions included the inspection tasks, work environment, and the tools available for use. Conditions such as temperatures and noise levels were not controlled but were recorded during inspections.

The AANC hangar is generally a low-noise environment, with few other concurrent activities. The hangar is clean. The floor is new asphalt. White drop-cloths were spread under the wheel wells and rear belly of the airplane to simulate lighter-colored concrete and/or painted floors.

2.6.2 Task and Equipment Conditions.

The inspectors were asked to finish each job card in the time that they considered normal and usual on similar jobs at their work location. ATA members of the steering committee assisted VIRP personnel in determining typical times expected for the selected job cards. The inspectors were assigned the inspections by being handed job cards that were similar to those used in the industry. Multiple copies of each job card were printed. For each job card, each inspector was given a clean copy and could make notes directly on the job card.

All inspectors received an identical briefing on the flight and maintenance history of the aircraft. The information was presented on videotape to ensure uniformity. The inspectors were requested to perform the inspections as if they were working as part of a team doing a D-level check on the aircraft. However, no repairs or destructive examination (drilling out rivets, removing or scraping paint, etc.) were to be made. That is, inspectors were asked to use their normal inspection procedures except where those procedures would leave marks or alter the nature of the flaw sites. This was done to keep the experiment conditions as fixed as possible from start to finish. Stickers were provided for marking components or areas of the aircraft, and inspectors were asked to mark flaws with these stickers.

A standard toolbox was furnished to each inspector. The toolbox contained a flashlight, dental-type mirror, adjustable angle mirror, 5x and 10x magnifiers, 6-inch scale (marked in mm and 1/100 of an inch), and a card of 28 numbered stickers (round, colored dots, about 0.75 inch in diameter). The stickers were used to mark areas called out as containing reportable discrepancies. Additional cards of numbered stickers were provided to the inspectors as needed. Portable fans, heaters, and area lighting were available for use, just as they would be in most facilities. Tasks were selected so that minimal scaffolding was required. Scaffolding was furnished—along with footstools. Ambient light levels and temperatures were measured at the time of each inspection.

3. IMPLEMENTATION OF BENCHMARK EXPERIMENT.

The complete activities for the Benchmark experiment were detailed ahead of time in a set of protocols that provided the briefing and debriefing materials and questionnaires and described monitor tasks and activities step by step. That is, the protocols provided a fully proceduralized job-performance aid for the monitors as well as the questionnaire forms to be completed with information from the inspectors.

Data collection consisted both of notes taken by monitors and videotaping of the inspections. Two monitors were used, one took notes and noted times while the other taped the inspection. These activities were alternated between the monitors during the various job cards performed by each inspector. Video tapes were used to validate questionable entries in the notes as well as to resolve obvious errors or omissions in the notes. Videotapes of the simulated lap-splice inspection were also used to separate and analyze search and decision behaviors (section 5.2.2).

3.1 SCHEDULES AND DURATION.

Two inspectors were scheduled per week. The experiment began on January 23, 1995. Two inspectors were observed that week and two more the next week. The final inspections were performed the week of March 20, 1995.

Each inspector was allowed 2 days to complete the preinspection questionnaires, the 10 inspection job cards, the post-inspection questionnaires, and the psychological tests (which took about 2 hours). The job cards were randomized in different orders for each inspector prior to the

arrival of the inspectors. Time considerations, especially near the end of the last shift, caused a slight change in the ordering of the job cards—so that, if the inspector was not going to be able to finish all the job cards, the same job card would not remain uncompleted for more than one inspector.

On Job Card 701, the first six inspectors inspected the large panel first, then inspected the small coupons. The monitors noticed that the inspectors were moving very slowly on this job card. Both discussion with inspectors (after the shifts were completed) and deduction on the part of the monitors led them to the conclusion that the large panel, with its fewer cracks, was leaving the inspectors wondering just what they were supposed to be looking for. Inspectors tended to speed up after they had seen the first large crack on the coupon set. Therefore, the monitors decided to move the large panels to the end of the row, so that the last six inspectors looked at the small coupons first and then inspected the large panel.

3.2 ENVIRONMENTAL CONDITIONS.

The first group of four inspectors were observed in late January and early February. The first two inspectors worked in cold conditions, although the weather began to moderate in February and, aside from occasional cold days, the temperature remained pleasant for the rest of the inspections into late March.

Lighting was somewhat typical of a hangar environment. Usually the light level was too low to measure with the available light meter (below 100 lux or 10 foot-candles). When the hangar doors were open, however, light levels were as high as 400 lux (40 foot-candles) for tasks on the outside of the aircraft and approached 1500 lux (150 foot-candles) for the inspection of the artificially cracked panels. Inspectors almost always used the furnished flashlight, providing at least 400 lux of illumination directly in the flashlight beam.

3.3 DATA COLLECTION.

Two primary types of data were collected. The primary data were the activities and accuracy of the inspection itself. The secondary data were characteristics of the inspectors, the environment, or the inspectors' behaviors that might have value in explaining the accuracy of results obtained.

3.3.1 Inspection Performance Data.

Each inspection started by recording the inspector code, date, time, job card number, hangar temperature, light level at the inspection site, and a description of the position of auxiliary lights as well as the starting location for the inspection.

Each inspection was recorded on a video tape that was labeled with the inspector code, date, time the tape was started and ended, and the job card numbers that were on the tape. The videotaping also contained the date and time recorded on the video. In addition to the videotape, another monitor recorded reportable flaws (squawks) and comments on a separate sheet.

The recording sheet served several purposes. First, as was explained to the inspectors taking part, the monitors recorded their findings so that the inspectors did not have to take the time to do so. Thus, no time was consumed by the inspectors writing, since accuracy of recording findings by the inspectors was not part of the experiment. The second purpose of the comment sheets was to record pertinent comments made by the inspectors (and their times) that were not part of a specific finding. Such comments included perceived condition of airplane, customary policy or procedures at the inspector's work location, comments about repair adequacy, etc. A third purpose of the comment sheets was to record monitors' comments about unusual or otherwise noteworthy actions, behaviors, or occurrences that could be germane to the inspection results. The monitors recorded the time of all comments made to correlate them with the videotape recordings.

The method of recording inspector findings was determined prior to beginning the experiment but evolved somewhat during the course of the observations. Originally, the monitors intended to record the inspectors' statements verbatim. However, this proved unwieldy because some inspectors were wordy. The monitors decided to record only essential data in abbreviated form—ensuring that time, flaw type, and location (body station and stringer/body-line) were recorded for each sticker. Flaw types were coded as cracks, corrosion, defective components (worn, broken, gouged, or missing), or dents. If an inspector thought an area needed to be cleaned better, this was also classified as a finding.

3.3.2 Inspector Characteristics Data.

Other factors of interest about visual inspectors in the Benchmark experiment were recorded. These factors were not, however, controlled or used as a basis of selection of the inspectors. These characteristics of the inspectors were recorded during the Benchmark experiment so that future studies can be compared to the Benchmark results. These characteristics are:

- Training
- Visual acuity
- Age
- Previous aircraft experience
- Education level
- Visual inspection experience
- Visual inspection experience by aircraft type
- Visual inspection training

In addition to these data, at the beginning of each shift inspectors were asked how well they had slept and what was their general physical, emotional, and mental condition. At the end of each shift, inspectors were asked what their physical condition was, whether they felt tired, and for some general ratings on attitude and attention as well as some questions on the effect of videotaping and the presence of the monitors on their work.

Information was also gathered at the end of each job card inspection. The inspectors were asked to rate their perceptions of the ease of tasks within the job cards, whole body exertion required to

perform the tasks, body part discomfort while performing the tasks, and their ratings of ease of physical access, ease of visual access, and comparability to typical shift conditions. Finally, the inspectors were asked how long since they had done this type of inspection on a B-737 and how long since they had done this type of inspection on any type of aircraft.

4. DESIGN AND IMPLEMENTATION OF FLASHLIGHT LENS EXPERIMENT.

For the flashlight experiment, sixteen of the eighteen lap-splice panels used in JC 701 were used. They were arranged in sets of four panels under two ambient lighting conditions. "Low" illumination was 90 lux (9 foot-candles) at the rivet level, while "high" illumination was 900 lux (90 foot-candles) at the rivet level. All panels were arranged so that the rivets were at mean (male) eye level of 60 inches (1.5 m) above the floor.

Twenty-four aircraft maintenance technicians, of whom twelve were qualified inspectors and twelve were mechanics or in other aircraft maintenance related jobs, took part in the experiment. They arrived in groups of six for 1-hour sessions per group. During the hour, each inspector was given 10 minutes to inspect each of the four sets of panels. Matched flashlights with original and light shaping diffuser lenses were used so that each subject inspected a different set of panels under all four combinations of flashlight lens and ambient lighting. Calibration of the rates of decrease of light output of the flashlights as batteries depleted in two runs over 2 days showed that batteries should be changed after two or three 1-hour sessions to keep the illuminance above 3000 lux. Subjects marked all findings on a sheet which reproduced the pattern of rivets on the panels. In the remaining two 10-minute periods of their hour, subjects performed a set of tests (near and far visual acuity, color vision, peripheral visual acuity) and completed a demographic questionnaire giving their age, training, experience, and use of corrective lenses.

The subjects were briefed to mark individual cracks on the right or left side of each rivet. The subjects' findings were compared to the known locations of cracks long enough to extend beyond the edge of the rivet head. Results of the analysis are given in section 5.2.3.

5. RESULTS.

In this section specific results from the Benchmark experiment and the flashlight lens experiment are presented. In presenting the results, the inspectors were randomly numbered 1 through 12 to preserve confidentiality.

5.1 JOB CARD TIMES.

Before looking at inspection results, we will look at the times taken by each inspector to complete the various job cards. The times (to the nearest 5 minutes) are given in table 1. The last column gives the average time taken for each job card, where the average is taken over all the inspectors. The bottom row shows the average time used by each inspector for a job card. In later sections the job card times will be compared to performance results. The purpose of this section is to look for patterns in times that could influence later comparisons.

TABLE 1. INSPECTION TIMES (MINUTES) BY JOB CARD AND INSPECTOR

JC	Insp 1	Insp 2	Insp 3	Insp 4	Insp 5	Insp 6	Insp 7	Insp 8	Insp 9	Insp 10	Insp 11	Insp 12	Avg.
501	85*	85	215	140	165	55	90	85	75	120	115	145	122
502	35	35	40	45	35	20	20	30	20	10	20	25	28
503	80	100	#	70	60	50	55	55	90	65	115	90	75
504	60	70	70	45	80	75	65	70	60	55	60	105	68
505	50	30	40	45	35	35	40	40	20	30	30	45	37
506	175	125	105	85	105	75	75	100	125	65	80	130	104
507	145	105	135	110	90	55	100	95	65	95	65	85	95
508/9	35	50	20	50	30	30	30	20	25	35	50	50	35
510	20	10	10	20	15	10	15	20	10	15	20	25	16
701	50	60	75	55	95	40	30	45	55	15	30	30	48
Avg.	82	67	79	67	71	45	52	56	55	51	59	73	

*Inspector 1 completed half of JC 501. Averages are based on doubling the time.
#Inspector 3 did not do JC 503. Marginal averages do not include this cell.

The average time per job card ranged from 16 minutes (JC 510) to 122 minutes (JC 501). Inspector 1 had 735 minutes (over 12 hours) of inspection and Inspector 3 had 710 minutes (almost 12 hours) of inspection. Both Inspectors 1 and 3 did not finish the inspections and it is estimated that approximately 90 minutes would have been needed for each to finish all the job cards. On the other hand, the fastest Inspector (6) completed all inspections in 450 minutes ($7^1/_2$ hours). Thus, across all job cards, the slowest inspection time was about 1.8 times that of the fastest (825 versus 450). The slowest time was approximately four times that of the fastest for certain job cards (JC 501 and JC 502). These ratios are quite typical of skilled operators performing well-practiced tasks.

Given the observed time differences in the job cards, are there systematic effects due to job card or to inspector? Or are the various inspector job card times showing random variation? To answer this question the logarithm[1] of the times were analyzed in a two factor (inspectors and job cards) analysis of variance. As expected, the time data clearly show both a job card effect (significance level, $p<0.001$) and an inspector effect ($p<0.001$).

Times for the first one or two job cards for a given inspector would be expected to be higher than normal if there was a "settling-in" period where inspectors became comfortable with the conditions. Also, one might expect an inspection time to reflect the number of calls being made. Therefore, we also analyzed the time data with respect to job card order and number of calls made in a given job card. The job card order was not a significant factor ($p=0.11$) in explaining the time variation, but the number of calls was a significant factor ($p<0.001$). Estimates from the

[1] Logarithms help normalize the data so that analysis of variance assumptions are met.

analysis indicate that, on average, inspectors would take about 25 additional seconds for each call they made. This estimate is consistent with the general process of removing a sticker, placing it at the called site, and speaking the call for the monitors to record.

An average of 24 calls were made per job card, but the number of calls varied substantially with inspectors as well as with job card. Much of the variation was due to different reporting styles, especially for calls of corrosion.

5.2 INSPECTION FLAW FINDINGS.

In this section we review the results of the inspections. Tables of flaws occurring on the Boeing 737 test bed are given. In some cases approximate crack lengths of the flaws are included. These are presented as background information. No attempt was made to analyze the implications of any particular miss on the ultimate safety of an aircraft flying with that particular flaw. There are a multitude of additional factors that would mitigate inspection misses with respect to ultimate safety, which are outside the scope of the analysis presented here. Suffice it to say, however, that the size of many of the cracks presented here are less than the size that aircraft manufacturers and airlines treat as detectable in establishing appropriate inspection schedules. Thus, the question in many cases should not be "Why did so many inspectors miss this crack?" but rather it should be "Why or how were a few inspectors able to detect such obscure cracks?"

5.2.1 EC Panels—Benchmark.

In this section we will present the results of the visual inspections performed on the manufactured lap-splice panels containing cracks of known length (JC 701). These panels were a subset of those used in an earlier eddy-current experiment [29]. The results of this portion of the Benchmark include the derivations of probability of detection curves that can be compared to previous eddy-current results.

Table 2 presents a summary of the detected cracks that were at least 0.050 inch in length as measured from the rivet shank. Cracks smaller than this would be almost entirely under the countersink rivet head and would not be visually detectable. Inspectors were asked to identify any cracks that they detected. Some of the rivets had pairs of cracks. That is, a crack emanated from the right and the left side of the same rivet. Cracks not called but located at a rivet with one that was called are listed in table 2 separately from the misses of a single occurring crack. The reasons for missing a crack located at a rivet with another crack may be related to the procedure employed by the individual inspector. This is best illustrated by Inspector 3 who, for the most part, only identified rivets containing cracks and did not specify the number or position of cracks observed at a particular rivet. The entries in table 2 are as if Inspector 3 called only a single crack at the rivets having two cracks.

False calls shown in table 2 are defined as the rivet sites for which the inspectors made a call and it is known that there was no crack. There was a total of 382 sites in the inspection for which such a call could be made. It should be noted that all the inspectors commented about the need for eddy-current verification or some other form of nondestructive inspection on their calls.

Therefore, the false calls represent the number of times that a nondestructive inspection follow-up would likely have found no cracks.

TABLE 2. SUMMARY OF CRACK DETECTION IN JC 701

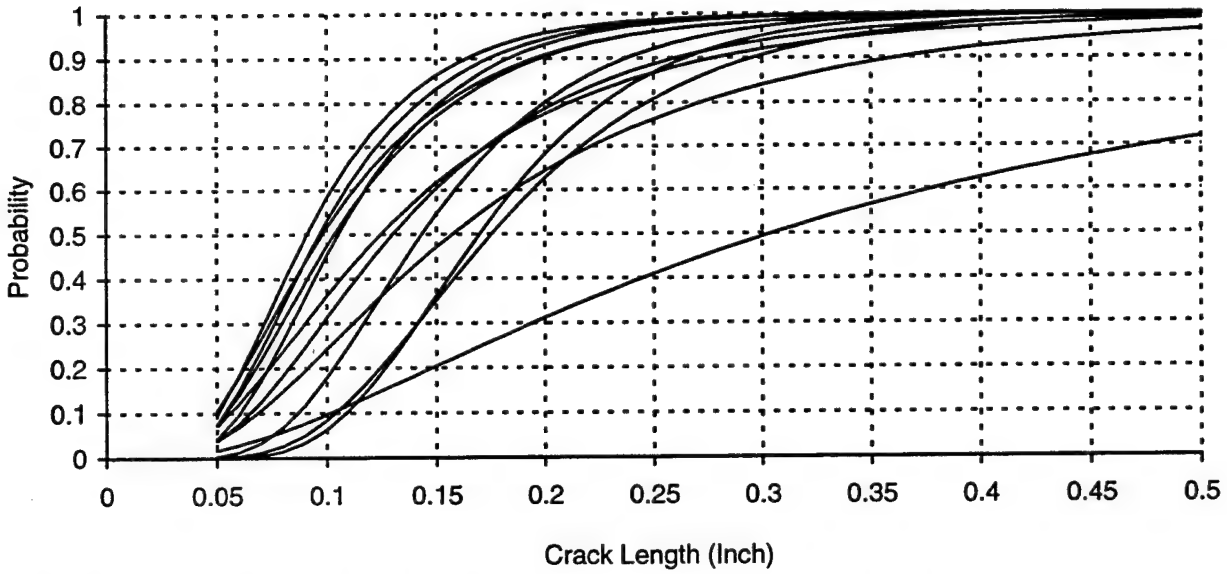
	Inspector											
Cracks (117 Total):	1	2	3[1]	4	5	6	7	8	9	10	11	12
Detected	71	22	26[1]	70	57	73	69	53	77	37	45	50
Missed	37	89	74[1]	45	55	41	45	56	35	74	66	63
Colocated Misses*	9	6	17[1]	2	5	3	3	8	5	6	6	4
"False Calls" (382 possible)	9	3	10	15	4	40	2	26	29	6	4	3

*Number of cracks not reported but located at a rivet where another crack was reported.
[1]Inspector 3 called rivet sites - did not identify individual cracks for the most part.

The summary given in table 2 illustrates substantial variation in the inspection results. Missing from table 2 is information concerning the size of the detected cracks. The size of the crack information is used to fit probability of detection (PoD) curves to each inspector. The probit methodology (discussed in [29]) was used to fit the curves shown in figure 1. The curves of figure 1 are terminated at 0.050 inch because lengths smaller than this would be under the countersunk rivet head and would not be visually detectable.

The same lap-splice test specimens were previously used in an eddy-current inspection reliability assessment [29]. In these results there was substantial inspector-to-inspector variability, but a typical eddy-current inspection achieved 90 percent detection rate around 0.100 inch.

In fitting the curves of figure 1 the colocated misses in table 2 were not included in the data. That is, the inspector was not given credit for finding the crack, but the crack also did not count as a miss. The effect of this assumption on the fitted probability of detection curves is given in table 3, where the 50^{th} percentile and 90^{th} percentile values of the fitted probability of detection curves are given. The effect is quite small for all but Inspector 8. The reason for the small influence is that the cracks that were not called but were located with other cracks had lengths that occurred in the middle to upper portion of the fitted PoD curve. This was not the case with Inspector 8. Two of the eight misses were extremely large cracks and therefore had a large impact on the estimated PoD curve. The two cracks were located on adjacent rivets and came together between the rivets.



Note: No penalty for not calling crack located at rivet with crack that was called. Crack lengths are measured from rivet shank, thus lengths less than 0.05 inch are not visible.

FIGURE 1. PROBABILITY OF DETECTION CURVES—JC 701 BY INSPECTOR

TABLE 3. PROBABILITY OF DETECTION CRACK SIZES

Inspector	50% Crack Length (Inch)		90% Crack Length (Inch)	
	Curve Fit Using All Cracks	Curve Fit Without Colocated Misses	Curve Fit Using All Cracks	Curve Fit Without Colocated Misses
1	0.10	0.10	0.20	0.17
2	0.32	0.30	0.95	0.91
3	0.18	0.17	0.29	0.26
4	0.10	0.10	0.20	0.20
5	0.13	0.12	0.31	0.28
6	0.10	0.10	0.19	0.19
7	0.11	0.11	0.19	0.18
8	0.15	0.13	0.46	0.26
9	0.09	0.09	0.18	0.16
10	0.18	0.18	0.29	0.29
11	0.16	0.16	0.37	0.36
12	0.14	0.14	0.24	0.24

Figure 2 shows the 0.90 probability of detection values (table 3 last column) versus time to complete JC 701. It is clear from the graph that over all the inspectors, the time taken to perform the task was not a good predictor of performance.

The inspectors participating in the Benchmark experiment are all active inspectors with substantial experience. The variation in the performance, as reflected in the curves of figure 1, can be used to gage the differences that would be statistically detectable in comparing different populations of inspectors. There are many ways to characterize the comparisons between two different (hypothesized) populations. Here, we consider the variation in the 0.90 probability of detection values given in table 3 (last column) and the differences that would have to be present in an additional group of 12 inspectors to deem them as having come from a different population.

The *t*-test is a standard statistical method for comparing two samples. We used a 90 percent confidence level *t*-test on the logarithm of the 0.90-PoD-crack size (written as $ln(a_{90})$). (The logarithm is an appropriate choice for transforming positive asymmetrically distributed variables.) The estimated median 0.90-PoD-crack length from the 12 inspectors is 0.258 inch. Assume another group of 12 inspectors were chosen for comparison and that group would have the same inspector-to-inspector variation in the response $ln(a_{90})$. The estimated a_{90} median from the second group would need to be smaller than 0.188 inch (a 27 percent drop) or larger than 0.353 inch (a 37 percent increase) to be considered statistically different from the original 12 at a 90 percent confidence level. Thus we see that fairly substantial differences have to be observed in two samples of size 12 from two different populations if they are to be judged different. Note that the underlying populations would have to be even further apart in performance levels to ensure a high probability of observing the above differences [30].

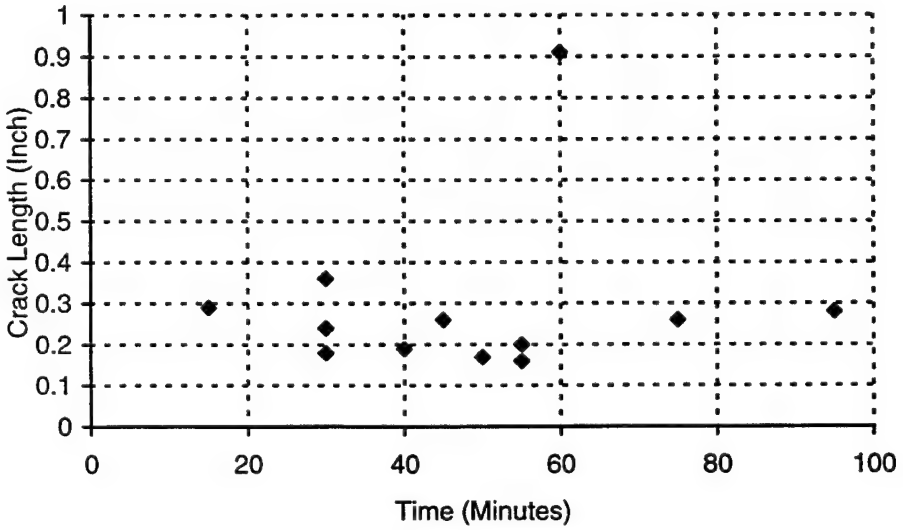


FIGURE 2. NINETY PERCENT DETECTION CRACK LENGTH VERSUS TIME ON JC 701

The PoD curves of figure 1 reflect several groupings of the inspectors. Although not labeled in figure 1, inspector numbers corresponding to the curves can be determined from the PoD values listed in table 3. Inspectors 1, 4, 6, 7, and 9 performed the best and Inspector 2 was the worst.

This is reflected both in the curves as well as with the total number of detections. However, the number of false calls for the upper group ranged from 2 for Inspector 7 to 40 for Inspector 6. Recall that false calls can reflect an inspector's decision that a particular spot should be inspected with nondestructive testing (NDT), such as eddy current. This false call variation suggests, however, that different decision criteria were being exercised by the various inspectors. This point is explored further in the next section where decision versus search behaviors are analyzed.

5.2.2 EC Panels—Benchmark Video Analysis.

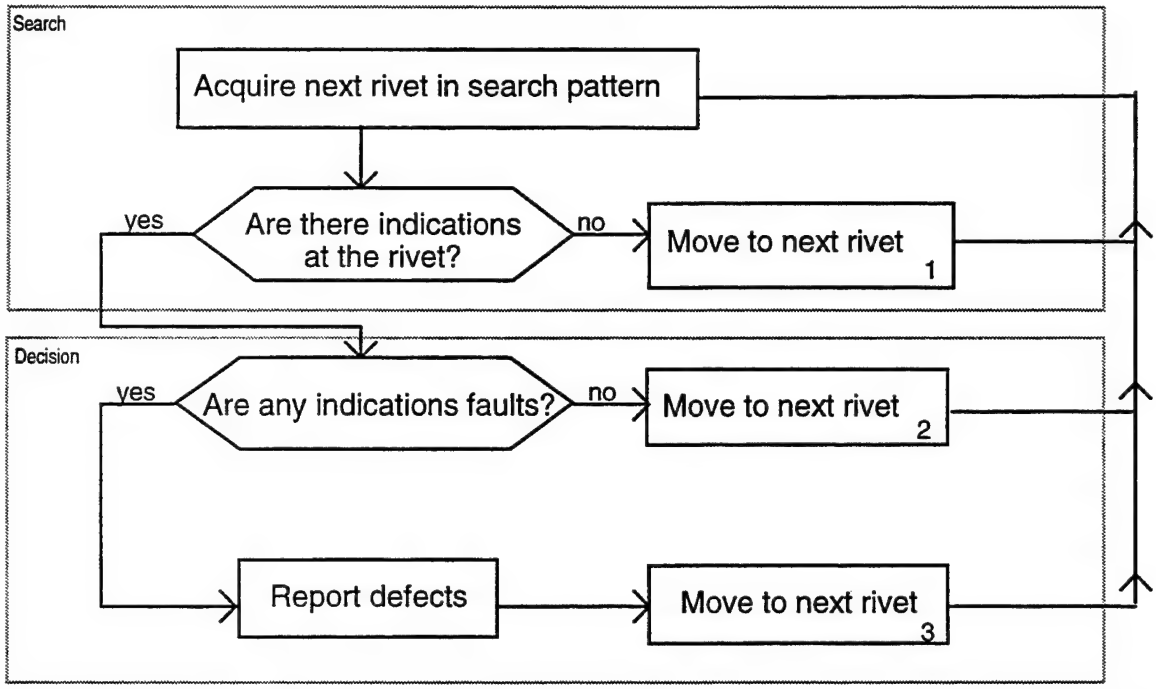
As discussed earlier, the inspection process is composed of two sequential subprocesses: search and decision. If we could characterize errors as "Search Errors" or "Decision Errors," it would help in understanding and concentrating interventions where they would be most effective [31]. To this end, we analyzed the videotapes of JC 701 inspections for information concerning the search and decision processes.

In search, the inspector covers the inspection area by a series of fixations separated by eye movements. The inspector will stop searching either because an indication is found or because he deems it no longer profitable to continue inspection, i.e., the area has been searched as thoroughly as desired but no indication has been found.

In decision, the indication located by the search process is examined more closely to determine whether it should, or should not, be called out. Usually this is done by a comparison of the indication with a standard for reporting of a defect.

We can thus characterize the inspection process as one of search followed by either a decision to stop searching or entry into a decision process, which may also include some element of search. The decision tree in figure 3 shows the different subtasks, the tests in each, and three possible outcomes. These numbered outcomes will have different final outcomes based on the actual state of the rivet as shown in the table at the bottom of figure 3.

In the Benchmark experiment, each inspector placed a numbered marker for each of his calls. Video recordings of all inspectors were taken so that it was possible to study how each inspector dealt with each rivet. The action of each inspector on each rivet was categorized according to the outcomes in figure 3. The flow chart reflects whether the inspector stops the search on a rivet and moves to the next rivet (outcome 1) or starts the decision process (outcomes 2 or 3). The decision step was indicated by a significant pause during which the flashlight and/or the inspector's head were moved in order to obtain additional views of a rivet. With the records of which rivets were marked and where the visible cracks (those which went beyond the diameter of the rivet head) were located, the search and decision process was reconstructed in this way. Note that search and decisions within a rivet site are not reflected in figure 3. An inspector could dismiss indications at a rivet site as being scratches, but in further search within the rivet decide that a crack was present. We have no way to quantify multiple search and decision loops within a rivet. Therefore, only the gross behaviors at the rivet level that could reasonably be characterized from the video tapes are reflected in the flow chart.



Outcome Number	Good Rivet	Defective Rivet
(1)	Correct search stop	Search failure: miss
(2)	Decision success: correct accept	Decision failure: miss
(3)	Decision failure: false alarm	Decision success: hit

FIGURE 3. INSPECTION DECISION TREE USED IN VIDEO ANALYSIS

Errors were classified using this methodology as due to a faulty search process (i.e., not even seeing an indication when one was known to be present) or due to a faulty decision process (i.e., misclassifying the indication once located). The importance of this analysis is that it allows countermeasures to be focused on the process where they are most needed. For example, quite different techniques are required for search training than for decision training [32]. An earlier study of manufacturing inspection of aircraft gas turbine roller bearings [31] used a similar analysis. This work helped to provide a more rational allocation of function between human and machine inspection and to design a highly-effective training system for inspectors [33].

Portions of each inspector's videotape were usable for analysis, except for Inspector 9. Table 4 shows the estimated probabilities for the different outcomes. It can be seen that the inspectors were poor at the search subtask, only locating 44 to 69 percent of the indications. They were highly variable in the decision process, with some inspectors perfect (Inspectors 5 and 7 both had 100 percent decision hits and no false alarms) and some very poor (Inspector 2 had few decision hits but not many false alarms while Inspector 10 had more decision hits but excessive false alarms). Thus, it appears that search interventions are required for all inspectors, but decision interventions are only required for a few inspectors. Most notably a decision intervention would be suggested for Inspector 2, whose performance on JC 701 was distinctly lower than the others.

That is, the analysis implies that the extreme PoD curve for Inspector 2 was not due primarily to search failures, but rather decision or classification failures.

TABLE 4. SEARCH AND DECISION PERFORMANCE MEASURES

Inspector Number	Prob. Correct Search Stop	Prob. Search Success	Prob. Decision Hit	Prob. Decision False Alarm	Overall Prob. (Hit)	Overall Prob. (False Alarm)
1	0.88	0.69	0.73	0.14	0.51	0.02
2	0.94	0.49	0.44	0.11	0.20	0.01
3	0.86	0.53	0.69	0.26	0.36	0.04
4	0.95	0.56	0.93	0.42	0.52	0.02
5	0.95	0.48	1.00	0.00	0.48	0.00
6	0.95	0.60	1.00	1.00	0.60	0.05
7	0.98	0.53	1.00	0.00	0.53	0.00
8	0.93	0.44	0.97	0.84	0.42	0.07
9	*	*	*	*	0.72	0.31
10	0.98	0.46	0.77	0.80	0.34	0.01
11	0.96	0.46	0.93	0.17	0.43	0.01
12	0.97	0.46	0.83	0.22	0.38	0.01

* indicates video tapes unusable.

5.2.3 EC Panels—Light Shaping Diffuser Flashlight Lens.

As presented in section 4, the eddy-current panels were used in a performance experiment comparing the use of flashlights with and without a light shaping diffuser lens. Because some rivets in these panels had two cracks, it was possible to measure performance reliability as "correct detection of a crack" or "correct detection of a rivet with a crack." False alarms could be expressed as a rate of rivets miscalled or sides of rivets miscalled. By using performance levels keyed to the rivets rather than the cracks at the rivets we acknowledge that some inspectors might approach the task as one of identifying rivets even though they were instructed to mark all cracks. All four performance measures for each combination of subject (24), flashlight lens (2), and ambient lighting (2) were examined.

Analysis of the results began by calculating correlations, r, between the performance measures. As expected, there were high correlations between the two measures of correct detections ($r = 0.906$) and between the two measures of false alarms ($r = 0.988$). The correlations were much smaller between the other pairs ($r = 0.145$, $r = 0.323$). Thus the choice of measures, i.e., defined by crack or by rivet, made little difference to the results.

Three-factor analysis of variance for each measure was performed. All four analyses showed significant differences between subjects, but no significant effects of flashlight lens, ambient lighting, or their interaction. Differences between subjects were further explored by correlating each of the four performance measures with the pretest and demographic variables. Only peripheral visual acuity gave significant correlation with correct detections. The correlation with correct crack detections over all flashlight and light conditions was $r = -0.61$, which was significant at a level $p < 0.01$. See figure 4 for a graph of detection rates versus the peripheral visual acuity sort time. Recall from section 4 that 12 of the subjects were qualified inspectors and 12 were mechanics or in other aircraft maintenance related jobs. These two groupings are also shown in figure 4.

Note that the correlation is negative. The peripheral visual acuity measure was the time to complete the peripheral acuity task, so that a negative correlation shows that subjects with <u>better</u> peripheral visual acuity (i.e., quicker times) were better at detecting defects. As peripheral visual acuity is known to predict search performance [34], this finding suggests that visual search is a limiting performance factor. Note that the video analysis of the Benchmark subjects also indicated that the search process was a performance limiting factor. Peripheral visual acuity is discussed in more detail in section 5.3.

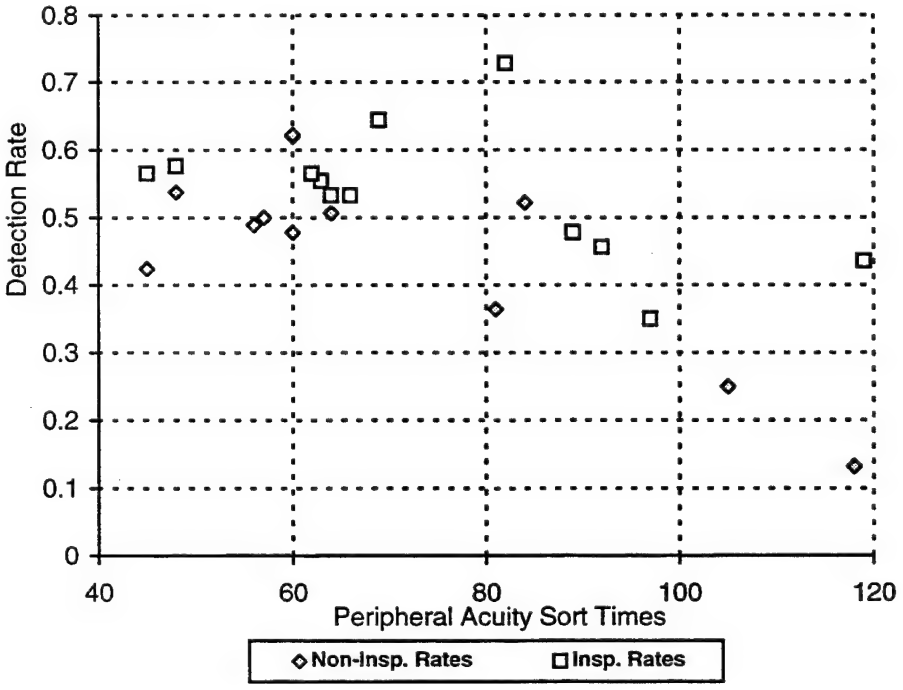


FIGURE 4. FLASHLIGHT EXPERIMENT DETECTION RATES VERSUS PERIPHERAL VISUAL ACUITY SORT TIMES

Significant differences were observed between subjects, indicating that the experimental design was sensitive enough to show individual differences in performance. To explore this factor further, the background information provided by each of the subjects was used to divide the

subjects into two populations: inspectors or noninspectors. There were 12 subjects in each category. An analysis was performed to test for differences between inspectors and non-inspectors on each of the measures. For the detection measures, these differences were significant at $p < 0.05$. There was no significant difference between the two populations on the false call measures. The mean performance of each group on each measure is shown in table 5 where it is evident that inspectors detected about 9 percent more flawed rivets (or 5 percent more cracks), a significant difference, and made 1 to 2 percent more false alarms not a significant difference.

TABLE 5. FLASHLIGHT EXPERIMENT PERFORMANCE BY JOB CLASSIFICATION

Group	Crack Detect Rate (%)	Rivet Detect Rate (%)	Crack False Alarm Rate (%)	Rivet False Alarm Rate (%)
Noninspectors	45	44	5	9
Inspectors	50	53	6	11

The mean performance under the four visual conditions is shown in table 6. Note that neither the ambient light difference nor the lens difference were statistically significant when compared to the inspector-to-inspector variation. The overall best level of performance was with the light shaping diffuser lens at the high ambient light conditions but this was not significantly different from the other conditions.

TABLE 6. FLASHLIGHT EXPERIMENT PERFORMANCE WITH LIGHTING AND FLASHLIGHT LENS CHANGES

Ambient Lighting	Flashlight Lens	Detections by Crack (%)	Detections by Rivet (%)	False Alarms by Crack (%)	False Alarms by Rivet (%)
Low (90 lux)	original	45	45	5	9
Low (90 lux)	light shaping	48	51	5	9
High (900 lux)	original	49	48	6	10
High (900 lux)	light shaping	49	51	6	11

The finding of no statistically significant difference between ambient lighting conditions was expected in that the illumination from the flashlight beam overpowered ambient illumination when the flashlight was held only a few inches from the rivet. Thus overall illumination was expected to be high enough for stable performance under high and low ambient lighting. However, the lack of a significant effect due to the flashlight lens was disappointing. While the enhanced flashlight did not increase total illumination, its effect was to provide better illumination uniformity. One reason for a significant difference not being detected may be due to restricting the test to detection of cracks in a bare metal panel at eye level. It is possible that different defects (e.g., corrosion) or different surfaces (e.g., painted metal, dirty surfaces, complex structures) would have revealed differences. The relatively small difference between

inspectors and noninspectors was statistically significant, so we know that if there is a flashlight lens effect it is not a large one. The conclusion is that under the conditions used the enhanced lens did not hinder but did not significantly enhance crack detection performance.

As a tie-in between the IAM inspectors of the flashlight experiment and the Benchmark experiment inspectors we compared their overall performances (percent cracks detected and crack false alarm rates). The IAM inspector group had a slightly higher average detection rate, but an average false alarm rate that was twice as high. Given the inspector-to-inspector variation that was present in both populations the average detection rates were not significantly different ($p = 0.32$), nor were the false alarm rates ($p = 0.11$). Thus there is no reason to believe that the two inspector groups represent different populations with respect to ability on this task.

5.2.4 Tie-Clip Inspections—On Aircraft.

In this section we will look at findings specific to JC 503. JC 503 included the inspection of tie-clips in the crown area of the fuselage interior. The tie-clip inspection is mandated by an Airworthiness Directive. The area covered by the job card contained a total of 98 tie-clips to be inspected. Of these, 24 were verified with eddy current to be cracked. Table 7 shows the number of detects called by each inspector as well as the number of tie-clips called where the eddy-current inspection did not verify the presence of cracks. (Inspector 3 did not do this job due to time constraints.)

TABLE 7. JC 503—TIE-CLIP INSPECTIONS

Inspector	1	2	4	5	6	7	8	9	10	11	12
Number found (of 24)	20	19	21	20	12	15	14	18	12	17	17
False calls made (of 74)	6	1	12	2	1	2	2	4	1	1	1
Inspection time (mins)	80	100	70	60	50	55	55	90	65	115	90

There is no significant correlation of the performance data of table 7 with the performance data in table 2 for JC 701. That is, knowing that one inspector did comparatively well on one of the job cards did not necessarily mean that he did relatively well on the other. This suggests that performance may be task specific.

Figure 5 shows the number of cracked tie-clips detected versus the time taken on inspection. The observed correlation is positive, but not significant (Spearman rank correlation = 0.44, $p = 0.10$, one-sided). It should be noted that the inspection for cracked tie-clips was only one task within a general inspection of the area containing the tie-clips.

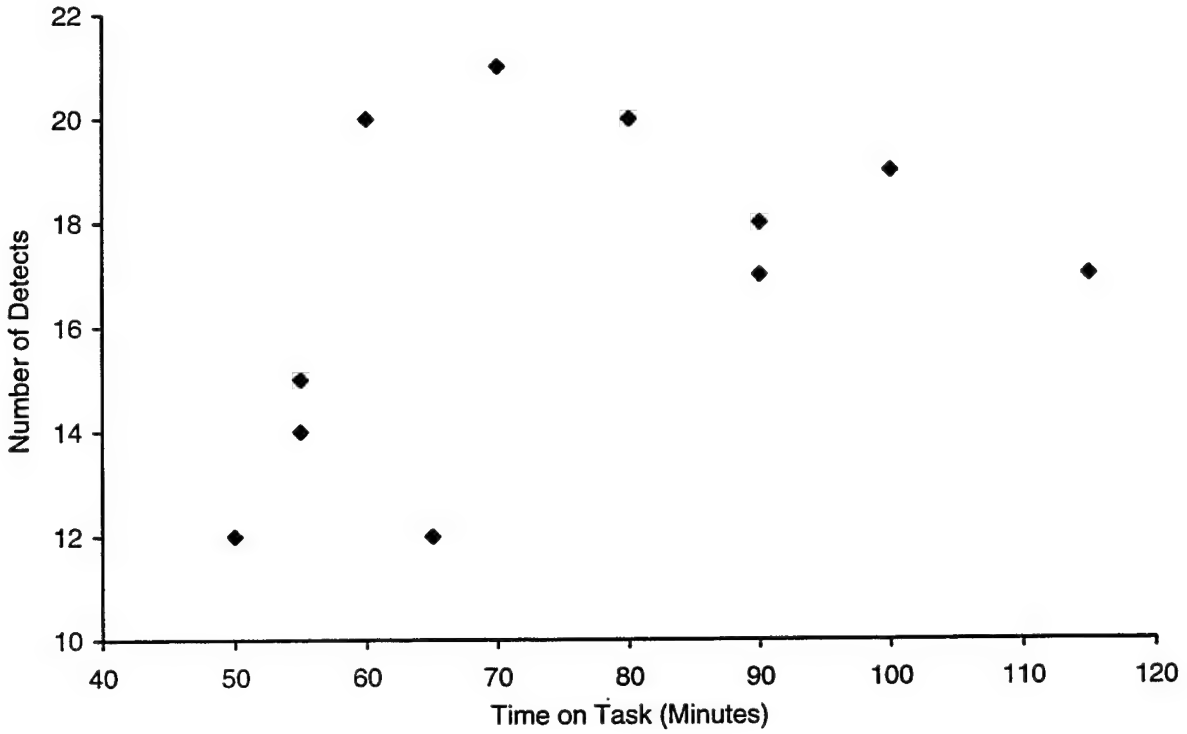


FIGURE 5. TIE-CLIP DETECTS VERSUS TIME TO COMPLETE JC 503

5.2.5 Flaws—On Aircraft.

In this section we focus on detection of flaws present in the Boeing test bed. All but one of the flaws were cracks. Table 8 gives the performance statistics on the 64 flaws. In many cases an approximate crack length is shown in the table. The flaws in table 8 are divided into three sections. The differences are discussed in the following subsections. Exact locations and descriptions are not given because of the continued use of the B-737 as a test bed.

5.2.5.1 Directed Inspections.

Two job cards contained Special Inspection instructions that identified typical cracks in the inspection area. The B-737 test bed contained the cracks as noted in the inspection instructions. The inspection outcomes for these two are given in table 8 as flaws 1 and 2. Flaw 1 was in a support structure and could be seen when looking straight up. This crack was missed by two of the twelve inspectors. One inspector called a possible crack on a nearby structure, but not the specific crack.

TABLE 8. AIRCRAFT FLAW DETECTION PERFORMANCE

Flaw	Approximate Length (Inch)	Ins 1	Ins 2	Ins 3	Ins 4	Ins 5	Ins 6	Ins 7	Ins 8	Ins 9	Ins 10	Ins 11	Ins 12	Total No. of Detects
1	1.0	Y	Y		Y	Y	Y	Y	Y	Y		Y	Y	10
2	8 to 11#	Y	Y		Y	Y							Y	5
3	4.5	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	12
4	0.7		Y		Y		Y							3
5					Y	Y		Y		Y				4
6	0.7	Y	Y	Y	Y			Y	Y		Y	Y	Y	9
7	1.5	Y	Y		Y	Y	Y	Y			Y			7
8	1.3		Y	Y	Y	Y		Y	Y	Y	Y	Y	Y	10
9	1.5		Y		Y	Y		Y	Y	Y	Y	Y	Y	9
10	1	Y	Y		Y	Y	Y	Y			Y		Y	8
11	.3	Y	Y	Y	Y	Y		Y			Y	Y	Y	9
12	.3	Y	Y	Y	Y	Y		Y			Y	Y	Y	9
13	0.8	Y	Y	Y	Y	Y	Y	Y	Y		Y	Y	Y	11
14	0.9	Y	Y	Y	Y	Y	Y	Y	Y		Y	Y	Y	11
15		NI*	Y	Y	Y	Y		Y	Y		Y	Y	Y	9 (/11)
16	3	Y	Y		Y	Y			Y				Y	6
17	3.3	NI		Y	Y	Y	Y	Y						5(/11)
18	0.8	NI	Y		Y	Y	Y	Y	Y				Y	7(/11)
19	1.5	NI		Y		Y		Y						3(/11)
20	0.2	NI	Y		Y								Y	3(/11)
21	1.2	NI		Y				Y			Y		Y	4(/11)
22	0.5	NI		Y				Y		Y	Y		Y	5(/11)
23	0.5	NI											Y	1(/11)
24	0.2	NI							Y					1(/11)
25	0.5	NI					Y							1(/11)
26	3.3	NI	Y		Y				Y			Y	Y	5(/11)
27	0.3	NI			Y									1(/11)
28	0.3	NI										Y		1(/11)
29	2	NI	Y	Y	Y	Y			Y		Y			6(/11)
30	0.9			Y										1
31	0.4			Y							Y	Y	Y	4
32	0.8												Y	1
33	0.5	NI											Y	1(/11)
34	0.5	Y										Y		2
35	0.5			Y			Y				Y	Y		4
36	0.3	NI		Y			Y	Y			Y	Y		5(/11)
37	0.5	NI		Y			Y							2(/11)
38	0.3	Y	Y		Y	Y								4
39	2.5	NI	Y			Y							Y	3(/11)
40	1.5	Y	Y			Y		Y					Y	5
41	0.4	Y	Y		Y	Y					Y			5
42	0.4		Y								Y			2

TABLE 8. AIRCRAFT FLAW DETECTION PERFORMANCE (CONTINUED)

Flaw	Approximate Length (Inch)	Ins 1	Ins 2	Ins 3	Ins 4	Ins 5	Ins 6	Ins 7	Ins 8	Ins 9	Ins 10	Ins 11	Ins 12	Total No. of Detects
43	1				Y								Y	2
44	0.5	Y												1
45	1.5	NI											Y	1(/11)
46	0.3		Y											1
47	1.6	Y	Y	Y		Y	Y		Y		Y	Y	Y	9
48	0.3	Y					Y	Y			Y		Y	5
49	1				Y									1
50	0.7							Y			Y	Y		3
51	0.6				Y						Y			2
52	0.3				Y	Y		Y	Y		Y	Y		6
53	0.8	Y	Y	Y			Y	Y		Y	Y		Y	8
54	0.2	Y												1
55	0.7		Y	Y	Y		Y	Y		Y		Y	Y	8
56	0.2	Y	Y					Y	Y		Y			5
57	1	Y	Y		Y	Y		Y	Y	Y	Y		Y	9
58	0.5	Y	Y			Y	Y							4
59											Y	Y		2
60	0.5		Y		Y	Y	Y						Y	5
61	0.8	Y					Y						Y	3
62	1		Y	Y		Y		Y	Y	Y	Y	Y	Y	9
63	0.3		Y			Y	Y		Y		Y			5
64	0.5						Y				Y			2
Totals		24 (/45)	34	22	31	29	22	28	19	10	31	22	34	

#Crack not visible for total length

*NI - Area with crack not inspected. The n in the margin total (/n) is the reduced sample size.

Flaw 2 was a crack in a bulkhead. The bulkhead is cracked at the point that the special inspection instructions indicates as a typical crack location. This crack runs along the edge of a stiffener on the forward side and is obscured. On the aft side of the bulkhead, the crack runs along the edge of a flat stiffener backing plate behind a fixture. The crack is thought to be large, but only portions (2 to 3 inches) of it are visible. It is almost impossible to detect with x-ray or eddy current without disassembly of the stiffener. Only five of the twelve inspectors reported this crack.

5.2.5.2 Apparent Flaws.

There were flaws in several places on the aircraft that were deemed as "easily detectable." These are shown in table 8 as flaws 3 through 15. Once these flaws are located, they are quite obvious and include cracks that are long and wide. Because of their obvious nature a miss within this set is considered more likely to be a search failure as opposed to a decision failure. Other flaws could arguably be added to this list, but these constitute the most obvious.

Crack 3 is in a location where the frame is broken in two by a crack that runs perpendicular to the length. Not only is the frame cracked but so is the fail-safe attachment (crack 5). Crack 3 is very wide and characterized by the inspectors as a "major break" or "frame severed." All inspectors saw and reported this crack. However, only three inspectors reported crack 4—another, clearly visible three-quarter-inch crack that was about five inches above the frame break on the top lip of the frame. Interpretation of this finding is difficult. It was clear that once in the area the attention of some inspectors was immediately drawn to the large crack 3. As a result the immediate surrounding area may have not been inspected as closely. However, at least one inspector moved on to other areas with the comment, "There's no sense in looking around here any more—the entire frame will be spliced anyhow." Similarly, only four of the twelve inspectors reported crack 5 at the bottom (skin side) of the broken frame. It is unclear whether the other inspectors did not look, did not see, or simply did not report the crack in the fail-safe since it would be replaced during the frame repair.

Crack 6 was located in another frame. Although this crack is not as large, it was gaped fairly wide and was about three-quarters of an inch long in a corroded area. (Corrosion finds are discussed in next section.) Nine of the inspectors reported this crack. The three inspectors that did not report the crack did report the accompanying corrosion (see table 9—corrosion 5).

Although physically different, cracks 11 and 12 were present in similar structure and were identified by most of the nine inspectors who detected the cracks as being in a "chronic" problem area in the aircraft. Cracks 13 and 14 were in symmetric structure (right side versus left side of aircraft) and were detected by all but one of the inspectors. Most of the inspectors also identified the area of cracks 13 and 14 as being "typically" cracked or in a "chronic" problem area. In both of the above pairs, the inspectors detecting one of the cracks also detected its twin. It was apparent that this is expected as many of the inspectors commented that the twin structure would be cracked once they made the first detection. Several of them stated that there would be cracks in these areas before they even looked.

There are likely many factors contributing to whether an inspector expects problems in an area and is therefore less likely to miss flaws. One likely factor is the familiarity of the inspector with the area on the particular model of aircraft. The one inspector that missed cracks 13 and 14 had recently inspected a B-737, but had not inspected this area on a B-737 in 2 years. This was the longest time gap for inspecting that area reported by any of the inspectors. One other inspector (who did make the detection) also reported not having inspected this area in about 2 years.

The time gap in inspecting the area containing flaws 11 and 12 is more varied among the inspectors making the detection as well as among the three that missed the flaws. One of the inspectors making the detection reported that he had never inspected this area of a B-737. Two other inspectors reported 3 and 4 years since having inspected this area on a B-737. The three inspectors who missed the flaws reported 6 months, 1 year, and 2 years since having inspected the area on a B-737. The above two examples show no clear relationship between the time gap in inspecting an area and the likelihood of missing a flaw.

Flaw 15 is not a crack, but rather a popped rivet. It is, however, an obvious flaw. It was missed by two of the inspectors and reported by nine of them. One inspector did not inspect the area containing the flaw due to time constraints.

In this group of apparent flaws, Inspector 2 detected all but one flaw and had one of the best performances. This is in contrast to his performance in JC 701. The video analysis of that job card indicated that Inspector 2 was failing in the decision process. Given his performance on the cracks present on the aircraft and the indication that the decision process was failing on JC 701, it is likely that the look of the cracks in the eddy-current panels was not what Inspector 2 expected a crack to look like.

5.2.5.3 Other Aircraft Flaws.

Flaws 16 through 64 in table 8 are cracks found by at least 1 of the 12 inspectors. This list does not encompass all of the calls made by the inspectors, but does include the majority of the cracks that were verified by eddy-current inspection. The crack lengths are based on visual information. There was no disassembly or laboratory characterization of the cracks beyond viewing them on the aircraft. Therefore the lengths should be considered as approximate lengths.

Is there a relationship between the crack lengths and the detection rates for this group of aircraft flaws? Figure 6 shows the detection rate versus flaw length for flaws 16 to 64 from table 8. Also shown in figure 6 are the detection rates for the flaws of JC 701. Probability of detection curves were fit to individual JC 701 data in section 5.2.1. The JC 701 cracks are tightly grouped in the scale presented in figure 6 but do exhibit a general increase in detection rates with crack length. On the other hand, the various cracks found on the aircraft have much more spread in length characteristics but do not show a strong relationship between detection rates and length of crack.

The detection rates for the aircraft cracks across the 12 inspectors have a small, but statistically significant, correlation with crack lengths. Using the same PoD curve fitting techniques as was used with the 701 data we obtain an estimate of a 50 percent detection crack (a_{50}) of 4.5 inches. Combining the data from all of the inspectors in JC 701 results in a a_{50} estimate of 0.13 inch. This difference is not unexpected. JC 701 was a specific task, inspecting for a single fault type, with specific areas to be inspected, all of which were consistent in structure. However, the aircraft flaws come from many areas and from different elements on the aircraft.

More telling than the difference in the a_{50} estimates is the differences in the statistical variation for those estimates. The 95 percent confidence interval for a_{50} in JC 701 is 0.12 to 0.14 inch. The same confidence level interval for the aircraft flaws is 2 to 72 inches. A confidence interval of this magnitude illustrates the inadequacy of trying to model the probability of detection solely as a function of crack length for such a varied data set. For a population of cracks from throughout the aircraft, crack length explains very little of the variation found in detection rates. This is not surprising in light of what is known about visual search in general, as discussed in section 1.3. The many factors that have been shown to affect probability of detection in search activities vary greatly across the various tasks. Thus, it is meaningful to fit probability of

detection curves as a function of crack length *only within specific inspection tasks and conditions of inspection.*

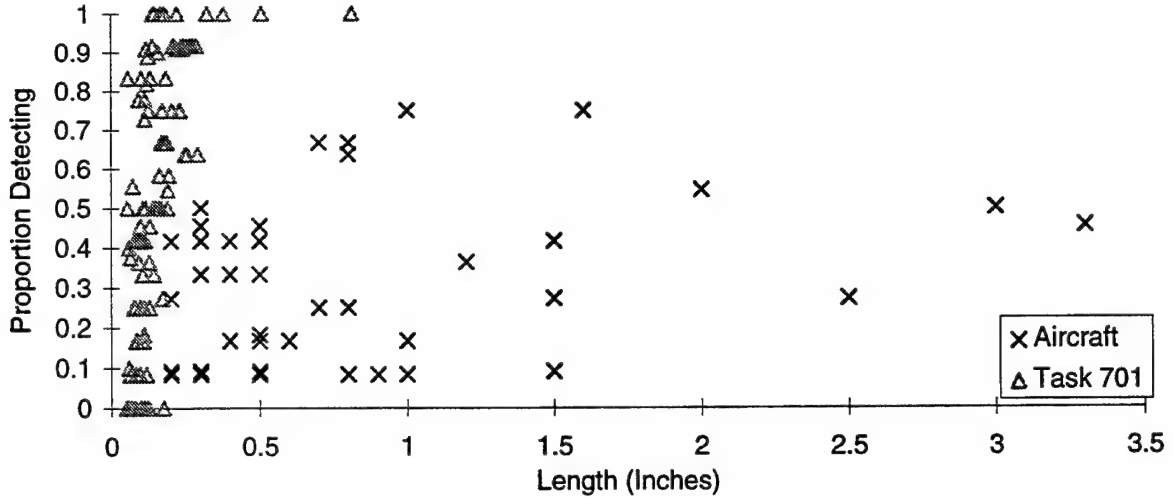


FIGURE 6. COMPARISON OF JC 701 AND AIRCRAFT FLAWS DETECTION RATES

5.2.6 Corrosion—On Aircraft.

Table 9 shows results for major corrosion areas on the B-737. There was an area of corrosion (corrosion area 1) that almost perforated the skin in the bilge. This corrosion could be seen from both inside and outside the airplane. Ten out of twelve inspectors reported this corrosion area; one other inspector called an immediately adjacent area. All 12 of the inspectors called the adjacent bulkhead attach angle corrosion (corrosion area 2). It is not clear whether the two inspectors did not call the specific area of corrosion in the skin because they assumed that other of their calls would initiate the proper repair.

TABLE 9. REPORT OF MAJOR CORROSION AREAS BY EACH INSPECTOR

Corrosion Area	INSPECTOR												*Total No. of Detects*
	1	2	3	4	5	6	7	8	9	10	11	12	
1	A^1	Y	Y	Y	Y	Y	Y	Y	Y	Y		Y	10
2	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	12
3	Y	Y	Y	Y	Y	Y	Y	Y	Y		Y	Y	11
4	Y	Y	Y	Y	Y	A^2	Y	A^2	A^2	A^2	Y	Y	8
5	Y		Y	Y	Y	Y	Y	Y	Y	Y		Y	10
6	Y	Y	Y	A^1	Y	A^1	Y	Y	Y	Y	Y	Y	10
7	Y	Y	A^3	Y	Y	Y	Y	Y	Y	Y	Y	Y	11
No. (of 7)	6	6	6	6	7	5	7	6	6	5	5	7	

A^1-Specific areas not called but adjacent areas were.
A^2-Skin immediately above lap-splice called as corroded.
A^3-Did not call specific area but inspector commented from earlier call he "would ask for skin to be opened."

The next two major corrosion areas involved lap-splices (corrosion areas 3 and 4). Corrosion area 3 consisted of two lap-splices that have been addressed by Service Bulletins. Eleven of the twelve inspectors called both lap-splices. The single inspector who did not call either lap-splice did call corrosion on adjacent skin areas. Corrosion area 4 was a lap-splice further down the side of the airplane. As can be seen in the table, eight of the inspectors called this lap-splice as containing corrosion. The other four inspectors did not call corrosion in this lap-splice, but they all reported corrosion on the upper skin immediately above the lap-splice. In making their calls, the inspectors were not always specific as to the nature of the visual clues. They did not get specific as to detecting corrosion residue, discoloration, or other signs, but rather just reported signs of corrosion.

Corrosion area 5 was on the frame containing crack 6. Both the inspectors that did not call corrosion did call the crack. Thus, all of the inspectors called at least one of the conditions, with seven of them calling both. It is not clear as to whether the inspectors making only one call did so assuming that any other flaw would be corrected during repair or whether the inspectors' focus on the one flaw resulted in a miss of another.

Corrosion areas 6 and 7 were externally viewed areas on the airplane. In limited areas of the inspection the bulging is enough that rivet heads have pulled through the skin and some cracking has occurred (corrosion area 6). All inspectors reported flaws in this area, but some reported cracks; some reported corrosion, and some reported both. The one inspector that did not call corrosion area 7 did not call anything in the specific location noted. However, signs of corrosion were reported in neighboring areas and the inspector said that he would "ask for skin to be opened."

Almost all of the inspectors commented that at the first signs of corrosion as bad as the indications of corrosion areas 6 and 7, their normal procedure would be to write up the whole area. Since a repair would encompass a skin replacement, there would be no reason to delineate individual flaw locations. In this sense, the major corrosion areas in table 9 would all have been adequately addressed in the working facilities of the inspectors. That is, a single callout would have initiated a repair action that would address all flaws in the area.

However, there are implications of a process in which a total delineation of flaws is abbreviated because the inspector believes the nature of the repair makes such a listing unnecessary. Specifically, the inspector has to be correct in his assumption of the nature of necessary repairs. The needed level of flaw delineation, as related to repair actions, is an item that should be addressed in airline maintenance facility's training programs for inspectors, if it is not already.

5.3 INSPECTOR DIFFERENCES.

In this section we summarize and discuss the background data collected on each inspector. We will also discuss associations of performance with these background variables as well as with data specific to each inspector, such as times on task.

Table 10 gives the minimum, maximum, and median values for several of the background characteristics. The "time since B-737 inspections-average all job cards" reflects the time gap in performing the specific inspections included in the 10 Benchmark job cards. It is calculated for each inspector by averaging their stated times since last B-737 inspection for each job card. The "time since any B-737 inspection" is the time since an inspector had performed any inspections on a B-737. Some inspectors had recent experience on a B-737, but not with the specific inspections included in the Benchmark experiment.

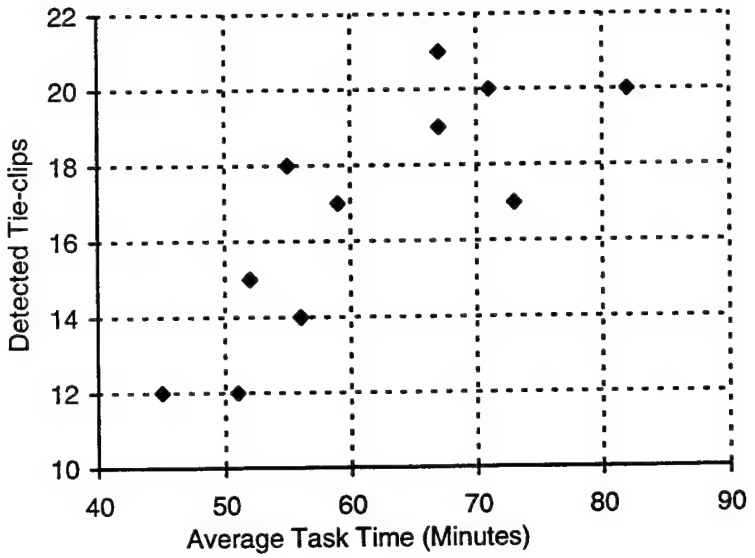
TABLE 10. INSPECTOR BACKGROUND SUMMARY

VARIABLE	MINIMUM	MAXIMUM	MEDIAN
Age (years)	32	50	38
Years in aviation	8	27	$14^1/_2$
Years as visual inspector	4	15	6.5
Time since any B-737 inspection	1 day	0.5 year	1 week
Time since B-737 inspections-average all job cards	2.5 weeks	3.5 years	8 months

As discussed in section 5.1 there was an inspector effect in the times taken to complete the various job cards. This means that over the various job cards included in the Benchmark experiment there was a general consistency in the approach of each inspector as reflected in the time needed to complete the jobs. In addition to the variables related to experience in inspection and experience on the Boeing 737, the directly measured times taken to perform each job are considered here as a reflection of inspector backgrounds.

The performance measures of JC 701 and the tie-clip crack calls of JC 503 did not correlate strongly with the times taken on each of those job cards. Here, we consider the correlation of performances with the times taken on all the job cards. There is little relationship between the average job card time and accuracy for JC 701. However for the tie-clip crack calls, there is a positive correlation (Spearman rank correlation = 0.77, $p<0.01$, one-sided), see figure 7. Thus, the inspectors who took more time (as measured over all the job cards) *tended* to be more accurate in their calls on the tie-clip task.

We can only guess why the association of tie-clip performance is stronger with overall times than it is with the time taken with the specific job card containing the tie-clip inspection. One contributing factor may be that individual job card times were only recorded to the nearest five minutes. Another contributing factor is suggested by looking at the top 3 tie-clip performances. They are three of the slowest times (1, 3, and 5 in ranks) in the average job card times over all inspections. They, however, are in the middle of the pack (5, 6, and 8 in ranks) with respect to the times specific to the tie-clip job card. The same three inspectors are ranked 1, 2, and 3 in the time (averaged over all job cards) since performing the various B-737 inspections. In general, these three inspectors were the most active in inspecting B-737's.



Note: One inspector (Inspector 3) did not do tie-clip JC 503.

FIGURE 7. TIE-CLIP DETECTS VERSUS AVERAGE JOB CARD TIME

Another fairly strong association was noted between tie-clip detection and the measure of peripheral acuity. Peripheral acuity was not measured directly, but rather the time to complete a card sorting task was measured. Subjects with better peripheral visual acuity have been shown to have quicker sorting times. The relative position of each inspector (detects as a percentage of detects from the top performer) is shown in figure 8 for both the tie-clip detects and the JC 701 detects.

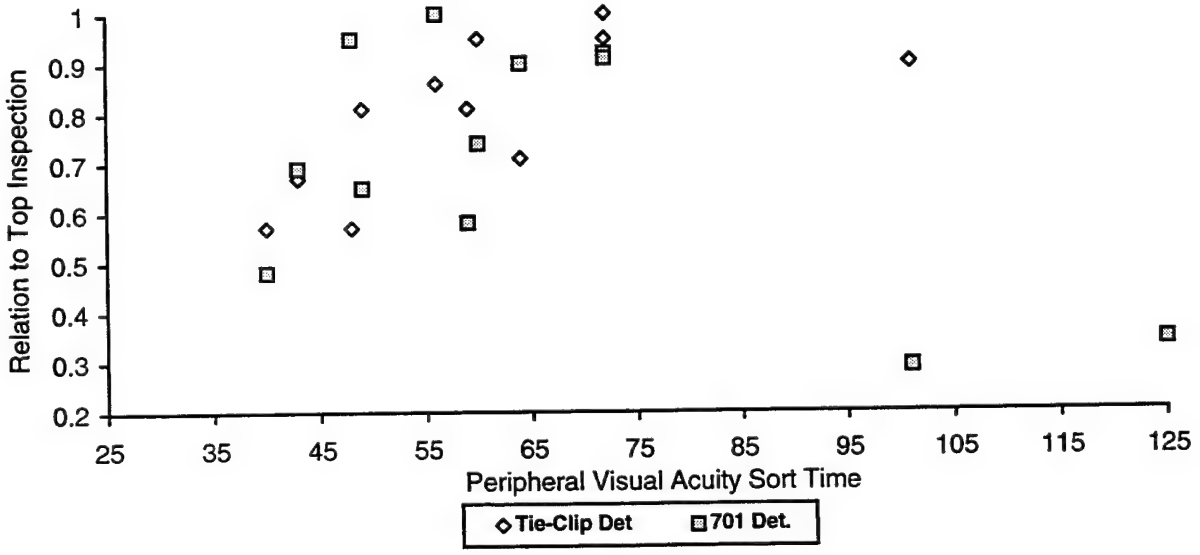


FIGURE 8. INSPECTOR POSITION VERSUS PERIPHERAL ACUITY

The peripheral acuity task time was shown to have a negative correlation with the detection rates of the inspectors in the flashlight lens experiment, as shown in figure 4. However, for the tie-clip job card in the Benchmark inspections there was a positive correlation, implying that better visual acuity was associated with worse performance. Although not statistically significant, the correlation of the JC 701 detects with peripheral acuity was negative as it was in the flashlight experiment data. At first glance these figures seem contradictory—better peripheral acuity is associated with better performance in some cases and with worse performance in others. However, the plots in figures 4 and 8 are more informative than the correlation coefficients. We note that with sort times between 40 and 80 seconds there seems to be a slight positive correlation with the various performance measures and the peripheral acuity sort times (figure 8). When the sort times go beyond 80 seconds, the negative correlation enters. In the IAM inspectors, there were enough subjects with greater than 80 seconds on their sort times that an overall negative correlation was implied. However among the Benchmark inspectors, there were only two whose sort times on the peripheral visual acuity task exceeded 80 seconds and one of them did not perform the tie-clip job card. Thus the tie-clip inspections exhibit a positive correlation of inspection results over the more limited range of sort times.

Age and the number of years in aviation were other inspector background factors associated with the performance measures. Of course, these two variables are correlated. In figure 9 the relative position of each inspector for the tie-clip inspections and the detections of the cracks in table 8 are shown as a function of the number of years in aviation. It is seen from figure 9 that the three inspectors with the most aviation experience were among the top performers on the aircraft tasks.

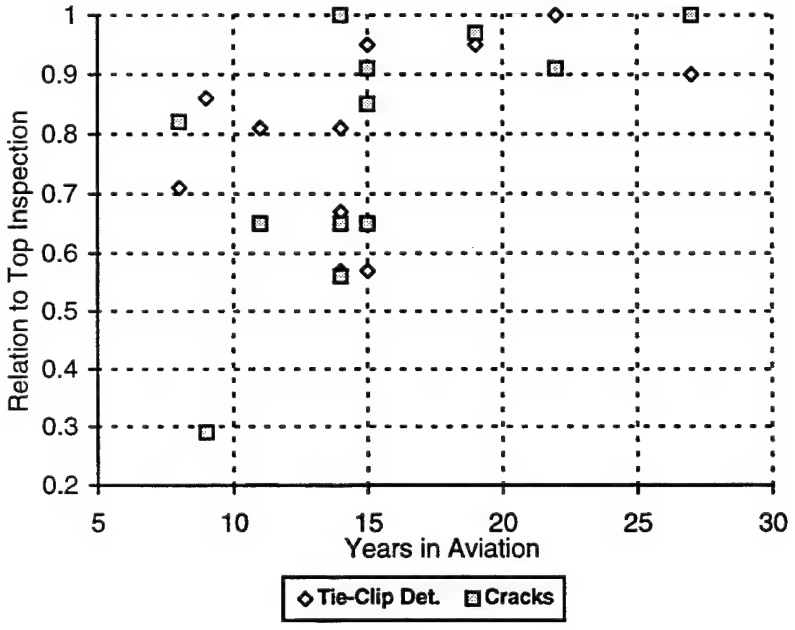


FIGURE 9. INSPECTOR POSITION VERSUS YEARS IN AVIATION

The inspector variables associated with performance levels that we discussed are time taken to perform all the job cards, peripheral visual acuity sort time, and years in aviation. In summary, worse peripheral visual acuity as measured by longer acuity task times was associated with lesser performance, but only in the extreme. Experience in the aviation field also was associated with better performance. There was a positive association of performance with more deliberate and unhurried behavior as measured by the overall time to perform the inspections.

5.4 JOB CARD RATINGS.

After each job card the inspector was asked to rate the job card as a single task on various numbers of standard scales as well as some developed specifically for aircraft inspection. The following aspects were chosen to cover difficulty, execution, physical discomfort, and physical and visual access. These were defined as follows:

- Task Difficulty. "How easy or difficult was this task?" Ten point scale (1 = Very Easy/ Highly Desirable, to 10 = Impossible).
- Whole Body Exertion. "How much effort did you exert in this task?" Ten point scale (0 = Nothing at All, to 10 = Extremely Strong/Almost Maximum).
- Body Part Discomfort. "How much discomfort/pain do you feel in each body part?" Five point scale (0 = None, to 5 = Intolerable) for each of 19 body parts. This was summed across body parts for initial analysis.
- Access to Workpoint. "The physical access to get to and from the workpoint was easy." Five point scale (1 = Strongly Agree, to 5 = Strongly Disagree).
- Access at Workpoint. "The physical access at the workpoint was easy." Same scale as the access to workpoint.
- Visibility of Workpoint. "It was easy to see the areas that were to be inspected." Same scale as previous two scales.
- Freedom from Distractions. "The surface areas to be inspected were free from excessive distractions with respect to the type of flaws being searched for." Same scale as previous three scales.

The analysis of this data considered agreement among the inspectors and the patterns of high or low values for each job card across inspectors. For each of the scale responses, an analysis of variance was performed with inspectors and job cards as factors and using the inspector—job card interaction as an error item. For all seven scales, the effects of both inspectors and job cards were highly significant ($p < 0.0001$). Inspectors were different in overall severity or leniency of their rating scores specific to a job card but overall they agreed on their relative ratings of each job card. In table 11, responses have been categorized as a "+" if the average response for that task was much higher than the average across all job cards, or a "-" if much lower.

TABLE 11. CATEGORIZATION OF INSPECTOR JOB CARD PERCEPTIONS

RATING SCALE	JOB CARD									
	501	502	503	504	505	506	507	508/9	510	701
Task difficulty	+	+	-	-	-	-	+	+	-	-
Whole body exertion	+	+		-	-		+	+	-	-
Body part discomfort	+	-		-	-		+	+	-	-
Access to workpoint				+	-	-		+		-
Access at workpoint		+		+	-	-		+		-
Visibility at workpoint		+		+	-	-		+		-
Freedom from distraction		+	-	+	-		+	+		-

Some important patterns emerge from this data. First, the three rating aspects of difficulty, exertion, and discomfort tended to give the same patterns across all job cards. Within these broad groupings, there were clearly some job cards which were considered "good" such as 505 (Rear Bilge Exterior) and 701 (Lap-Splice Panels) and some as "bad" such as 508/9 (Rear Bulkhead Y Ring) or 502 (Main Landing Gear Support). Between these were job cards with low difficulty but difficult access such as 504 (Galley Doors, Interior) or with high difficulty but neutral access such as 501 (Midsection Floor Beams).

We conclude that inspectors can provide useful data on their perceptions of the different tasks. Also the design of the experiment was successful in achieving a mix of high and low levels of difficulty and access across the set of all tasks.

5.5 OBSERVATIONS OF INSPECTION TECHNIQUES.

The monitors recorded the general behavior and techniques that were used by the various inspectors in their personal log books and in comments on the recording sheets. These observations were supplemented by casual discussions with the inspectors. Techniques that were prevalent in the inspectors' facilities and the inspectors' personal experiences were included in the discussions.

5.5.1 Job Cards and Supporting Materials.

As part of the Benchmark experiment each inspector was given the set of job cards that he was expected to complete during the shift. All job cards contained instructions for general inspections of an area along with instructions for detailed inspections within the area. Several characteristic techniques for using the job cards were displayed by the inspectors. All of the inspectors glanced over the set of job cards, at the beginning of the shift, to see what they were expected to do. At the beginning of each job card several different behaviors were observed.

Inspectors would generally follow one of these behaviors.

1. Look at the job card just before beginning, lay it aside, and not refer to it again.

2. Look at the job card once in the area to be inspected, lay it nearby, and possibly refer to the job card during the inspection but at least review the card at the end of inspection to verify the various elements had been completed.

3. Read the job card closely at the inspection site, comparing drawings with the structure he saw in front of him, then refer back to the job card throughout the inspection.

Although each inspector could, in general, be classified according to one of the above behaviors, some of them exhibited different behaviors for different job cards. Of course, factors such as the length and complexity of the tasks within the specific job card, as well as familiarity with the tasks, would affect how each job card was used. One inspector, who tended to follow behavior 1, announced his completion of one job card after a relatively short time. The incredulous response of one of the monitors, led the inspector to go back to the job card and review it. He then realized that he had performed the detailed inspections called for in the job card but had not inspected the complete area given for the general inspection. The inspector returned to the task and completed it, making three additional calls.

One inspector, who made frequent use of the job cards during inspections, remarked in one particular job card that the diagram showing a typical crack was responsible for him locating the crack. The crack (table 8, crack 2) is in the location specified in the job card drawing but is obscured by various elements on the aircraft and was detected by only five of the twelve inspectors. This particular inspector noted that he was led to a more intensive look in that area because of the diagram.

The above instances illustrate undesirable outcomes (areas not inspected) resulting from an indifferent or casual use of job cards as well as desirable outcomes (finding of a crack) resulting from a diligent and studious use of the job card. Unfortunately, as we try to break down the inspectors into categories of job card usage, we end up with small numbers of inspectors in each category. In light of the amount of variation in the inspection results, these small numbers do not allow statistically significant generalizations to the airline maintenance inspector population as a whole. However, these observations on the use of job cards are consistent with expectations and with previous research into work practices.

5.5.2 Systematic Search.

Some inspectors very clearly and openly used a standardized (for them) search pattern to ensure that they covered all of the area. These inspectors were able to describe the pattern that they were using. In certain cases, the inspector performed a systematic search in an area looking specifically for one type of defect (e.g., cracks) and would then search looking for another type of defect (e.g., corrosion).

Systematic search strategies, however, do not guarantee that no area will be missed. For example, some inspectors would proceed one bay (area between structural frame members) at a time—yet they would skip around within this restricted area. This effect appeared to be most common when the inspector's attention was drawn to a more obvious defect in the area. The inspector would mark and describe the flaw, then forget where he had been in the search process.

5.5.3 Knowledge of Chronic Problem Areas.

There appeared to be a real advantage for inspectors who were familiar with the model of aircraft that they were inspecting, especially when this familiarity included knowledge of the chronic or typical problems for the model. Cracks 11 and 12 in table 8 illustrate the point. These cracks were reported as a chronic problem by most of the nine inspectors that found them. Some of the inspectors even reported that their facility had produced a job card directing an inspection specifically to check for these cracks.

More than one inspector noted that they, and their colleagues, often approached specific areas in a confirmatory mode. That is, they were dealing with a chronic problem area and the intention of the inspection was to confirm or verify the presence of a flaw. It is likely that the expectation of a flaw enhances its detection, but this type of familiarity is a mixed blessing. Inspectors occasionally became so focused on finding known flaws or the special detailed inspections that they missed nearby flaws. For example, we watched one inspector search for a crack that he knew from past experience was at a specific location. He was so focused that he missed a clearly visible 1.5-inches crack only 3 inches above that location. (He even went and got a magnifying glass to search for the crack he knew had to be there.) We observed several more instances of this "tunnel vision," although not always where there were flaws in the immediate area.

On balance, however, intimate knowledge of the chronic or typical problem areas of a particular aircraft model is more help than not. In casual questioning and discussion with the inspectors, the monitors found that the inspectors relied first on personal experience and discussion with other inspectors to gain this knowledge and second, on the chronic problem or "bad actor" lists provided by their employer. It is not clear if the inspectors assimilate to the same degree the reporting of cracks found by other airlines. General reporting of such finds is often contained in background sections of Service Bulletins and is also implied by the inclusion of a typical crack in Service Bulletin drawings. This information is intended to provide this kind of familiarity with chronic problems across airlines for the entire fleet. However, this information is far enough removed from the personal experience of the inspector that it perhaps does not have the same impact, especially considering that some inspectors used the job card to establish the inspection task but they did not use it as a reference.

There may be several reasons that the job cards (and the information they reflect) do not appear to be fulfilling a function as vicarious experience. One reason is that they are often difficult to understand and apply. This is especially true of drawings. As an example, the special detailed inspection instructions for the main landing gear wheel well show key elements in a construction style drawing (see figure 10). However, the view is from a perspective that cannot be seen in the assembled aircraft. The inspector must mentally reorient the drawing and mentally superimpose

other aircraft structure to match what he can see. As decades of research have shown, humans do this sort of mental task very poorly, with a great deal of effort, and often make mistakes doing so.

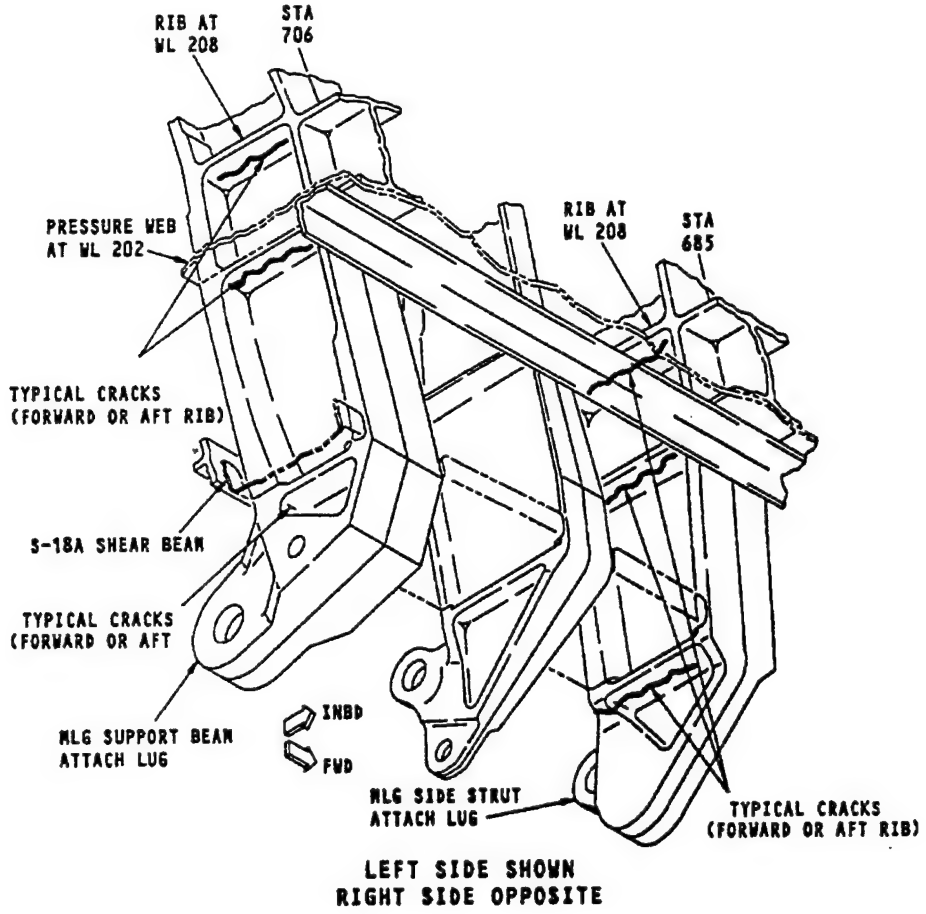


FIGURE 10. INSPECTION DETAIL FROM SERVICE BULLETIN DIAGRAM

5.5.4 Definitions of Boundaries for Inspections.

The monitors observed that when an inspection was stated to be from BS639 to BS727 (for example) different inspectors interpreted this information in different ways. Some inspectors would inspect from the aft face of the frame at BS639 to the forward face of the frame at BS727, ignoring the edge and other face of the frame. Other inspectors seemed to draw an imaginary line down the center of the edge of the boundary frames and inspect that edge and the aft and forward faces, respectively. Still other inspectors would examine the entire frame, forward and aft faces and edge on each of the boundary frames.

A procedural question related to what part of the boundary element should be included is the interpretation of boundary information as meaning "up to, but not including." That is, some inspectors would inspect an area defined as "from BS277 to BS540 and Stringer 4L to Stringer

17L" and inspect everything inside the boundaries but not inspect the frames and stringers cited as boundaries. Although the monitors casually questioned the inspectors, it was never clear as to whether these differences were due to differences in interpretation at different facilities or whether it was individual differences in interpretation.

It should be noted that to insure adequate inspection coverage and to avoid misinterpreting job card boundaries, inspectors often extended the inspection area to encompass adjacent areas. However, several of the inspectors noted that, in the past, they have been questioned by their supervisors why they are reporting discrepancies in areas that they were not specifically told to inspect. This observation and the difficulties with diagrams mentioned above reinforce the need to apply good human factors design techniques to job card design [35].

5.5.5 Appeal to NDI Instrumentation.

Although false alarm rates were calculated for the verifiable cracks of Job Card 701 and the tie-clips of Job Card 503, these rates overestimate the number of ultimate false calls as measured by repair actions taken that are not necessary. The reason that this number is an overestimation is that, frequently, the inspectors either said, "I would ask for an eddy-current check on this one." or they said, "I would grab an eddy-current machine and check this one out." In these cases, the stickers indicated an area that the inspectors are deferring decision making to other than visual means. Thus, they were calls for further examination and would not necessarily lead to needless repair or additional down time.

5.6 COMPARISON TO OTHER STUDIES.

A study of field data gathered over three years on the results of inspections of transport aircraft operated by Japanese airlines concluded that prior information, inspection distance, surface condition, and crack origin influenced results [26]. We briefly consider these points as they are related to the Benchmark inspections and our observations of those inspections.

Endoh and his colleagues [26] analyzed various factors by constructing normalized cumulative functions with respect to crack length for each level of a factor. Thus, for example, to look at the effect of prior information the field data were separated into two groups. The groups were those finds where prior information was and was not noted. The graphs of the normalized cumulative count as a function of crack length for each group were then compared. The resultant curves have the look of probability of detection curves, the ordinate being crack length and the abscissa being a proportion (0 to 1). They are, however, not PoD curves.

Probability of detection is only one factor of several that could give rise to differences in these curves. Another factor that could explain observed differences would be the underlying population of cracks for each characteristic. For example, directed inspections in a specific area might be carried out sooner and more often than the general inspections of another area. Therefore, the population of crack lengths at the time of inspection is naturally smaller in the directed inspections than in the undirected inspection and with no differences in probability of detection this would still be reflected in the crack sizes of reported field data.

In the current study inspection tasks were constant and differences are reflected by performance levels as measured over the 12 inspectors. With this introduction we consider the four factors identified in reference 27.

- *Prior Information*. We noted in earlier discussions that specific cracks were reported as being in chronic problem areas by the inspectors. Those inspectors expecting to find cracks found them when they were there. If expectations were high enough, and locations specific enough, the inspectors could be said to be operating in a confirmatory mode as opposed to a search mode.

- *Inspection Distance*. There was no specific control on inspection distance in the Benchmark inspections. Most of the inspectors exhibited inspection behaviors where they went through great pains to get their eyes close to the area being inspected. The monitors noted one exception. One inspector seemed to be less discomfort tolerant than the others. He related having devised an inspection method for a task in his facility that allowed him to inspect without having to bend over. The result was that the inspection would be done more removed from the area being searched. This same attitude and approach was observed in the Benchmark inspections done by this inspector. This particular inspector was the poorest performer on finding cracks on the aircraft. It is believed that this was due, in part, to not getting close to the inspection.

- *Surface Condition*. The Japanese reported that the detected crack lengths on dirty surfaces were longer than on clean surfaces. A common comment for all the Benchmark inspectors was that certain areas were in need of cleaning before an effective inspection could be done. We have no data on whether there is a difference in detectability for dirty versus clean surfaces, but it is clear that the inspectors expected the surfaces to be clean for them to do an effective inspection. They all reported being able to demand cleanings prior to inspection.

- *Crack Origin*. The Japanese data showed no differences between cracks emanating from fastener holes and those from edges. However, they reported a difference between these two categories and the other category. In general, crack length explains very little of the variation seen in detection rates for cracks from all over the aircraft (see discussion in section 5.2.5), but the data have not been fully analyzed with respect to other aspects of crack morphology.

6. SUMMARY AND CONCLUSIONS.

There are four broad areas related to visual inspection reliability that will be discussed and summarized here: quantification of inspection reliability, search and decision aspects of visual inspection, usage of job cards within inspections, and inspector-specific factors affecting visual inspection performance.

It should be re-emphasized that the scope of this program included inspection performance on a transport aircraft. As such, we have looked at inspection performance with respect to naturally

occurring flaws found on the aircraft. That is, the flaws arise from the stresses and the use conditions of the aircraft. The individual flaws have not been analyzed with respect to ultimate safety of the aircraft. Any given flaw may be well within the size assumed detectable by the maintenance and inspection programs. We are not raising alarms concerning low detection rates on any particular flaws or set of flaws but rather looking for implications about the inspection process.

It should be emphasized that this program gathered a lot of data related to inspection behaviors. The factors discussed here represent a top level look at that data. We expect that additional analysis of the data, including the videotapes of the inspections, is likely to yield additional insights into the inspection process.

6.1 PROBABILITY OF DETECTION.

Inspection programs are set using damage tolerance ideas integrated with probability of detection for specific types of flaws. The crack characteristic often used to express a probability of detection is crack length. In this program we included a set of lap-splice panels with known cracks from beneath rivets to estimate the probability of detection curves as a function of crack length for each inspector.

The probability of detection curves estimated for each inspector exhibited substantial inspector-to-inspector variation. Crack lengths at which a 90 percent detection rate is achieved ranged from 0.16 to 0.91 inch across the 12 inspectors. False call rates differed substantially even with comparable performance in detections. False calls in this context indicate that the inspector felt that the area required further checking with nondestructive testing.

One practical implication of the PoD variability is that any experimental program comparing two populations of inspectors (with a modest number of inspectors from each population) will be reliable in detecting only large differences in population means. More exact quantification of sample size requirements can be found in standard statistical texts and in reference 30.

For a population of cracks taken from many areas of the aircraft and from many different types of structure, crack length explains very little of the variation found in detection rates. Thus, it is meaningful to fit probability of detection curves as a function of crack length only within specific inspection tasks and conditions of inspection.

Quantification of performance for individual inspectors is task dependent. That is, knowing that an inspector does comparatively well on one task does not necessarily mean that he will do well on another task. This was indicated by the lack of high correlation in the performance levels of various tasks.

The Benchmark experiment reported here was successful in achieving a mix of high and low levels of difficulty and access across tasks. This was evident from the inspectors' evaluations of each of the job cards.

6.2 SEARCH AND DECISION.

The visual inspection process includes components of search as well as of decision. Video analysis of inspectors' behaviors on a task of finding cracks from beneath rivets indicates that the inspection process could be improved with search interventions for all of the inspectors. A small number of the inspectors would also benefit from decision interventions.

Although the evidence indicates that improvements in the search process for most inspectors is warranted, there is limited research available on search strategies specific to aircraft inspection. It is likely that good search strategies are task dependent.

One aspect of the search process handled differently by different inspectors was that of curtailing a search due to finds. For example, some inspectors relied on their belief of a likely repair action to dismiss further inspection once something major had been found. They assumed that a repair necessary for one flaw would remove other flaws if they were present. Although we had no basis to judge the correctness of this assumption when it was made during the Benchmark, this practice does raise some issues. Are there multiple repair actions that can be taken with respect to a given call? Is the inspector sufficiently knowledgeable about possible repairs or could he be making an erroneous assumption about the nature of a repair? Is the inspector calling the most significant damage in an area? These questions (and their answers) take on even more significance if the repair technician or even a different inspector conduct inspections after the repair. The bottom line is that maintenance facilities need to ensure that repair technicians actually repair everything in an area whether it has been explicitly called or not.

The worse spots of corrosion were called by all inspectors. Typically corrosion calls cover broad areas and inspectors rely on the repair process to address all manifestations of the corrosion.

6.3 JOB CARD USAGE.

There was not a consistent use of job cards between inspectors. Some inspectors used the job card only to establish the boundaries of the inspection task. The job card often contained figures giving locations of likely cracks and therefore could be a valuable reference. In one instance of a crack being present at the exact location shown in the job card, five inspectors detected it and seven inspectors did not. The misses are attributed, in part, to poor use of the job card. Reference to the job card during the inspection, or at the very least using it as a final checklist to verify a completed inspection would likely lead to better inspection performance.

Job cards are not without their problems. Typical diagrams of inspection areas often do not look at all like the area the inspector will be looking at on the aircraft. This can be due to not showing the structure as it would be viewed by the inspector. The diagrams often do not contain surrounding structure that may partially obscure the elements being presented. They may also give an orientation specific to only one side of the aircraft, expecting the inspector to be able to mentally reorient the whole diagram. These problems could be helped by the inclusion of photographs of inspection areas as they will appear and better graphical design showing not only the structure to be inspected but also, through appropriate shading, the structure that is obstructing a clear view.

Job cards also should be clear in defining the boundaries of an inspection. We observed that when an inspection was stated to be from BS639 to BS727 (for example) different inspectors interpreted this in different ways. Some inspectors would inspect from the aft face of the frame at BS639 to the forward face of the frame at BS727, ignoring the edge and other frame faces. Other inspectors would examine the entire frame, forward and aft faces and edge on each of the boundary frames.

6.4 VISUAL INSPECTION PERFORMANCE FACTORS.

The inspectors in the Benchmark program were observed on 10 different job cards. The job cards reflected variations in accessibility and in visual complexity. The times taken for each job card reflected both a job card effect as well as an inspector effect. That is, inspectors that were quicker in their inspections tended to be quicker across all the job cards. The time taken for a specific job card, such as the tie-clip inspections, did not exhibit a strong association with performance in that task. However, the inspectors who took more time (as measured over all the job cards) *tended* to be more accurate in their calls on the tie-clip task.

There were several background factors that were associated with performance. Peripheral visual acuity (as reflected in times to complete a controlled search task) and aviation background were associated with performance levels.

The extreme high values of the peripheral visual acuity times (less peripheral acuity) were associated with a drop in performance levels. However, over a more limited range of the peripheral acuity time, an increase in performance was noted. The observed association of performance with peripheral acuity is consistent with the observation that visual search is a factor limiting performance.

There was no clear relationship between recent experience in a specific inspection area and performance in that area. However, there was an association with aviation background and performance levels. The more experienced inspectors, as measured by aviation background (not inspection background), also performed better.

An issue related to the overall experience level is the expectation of finding cracks in specific areas. The Benchmark experiment included areas containing cracks that were expected by many of the inspectors, but were missed by others. The inspectors expecting to find these cracks looked directly at the area and confirmed their presence. The inspectors missing the cracks were observed to search the area (as evidenced by movement of the flashlight beam), but failed in the search process.

The down side of expecting to find a flaw in a specific area is a tendency to focus on that area and possibly alter search patterns in adjacent areas. This behavior was observed both when the expected flaw was present and when it was not present. In the latter case, the inspector devoted substantial time to a thorough inspection in the area he believed would be flawed and neglected immediately adjacent areas where flaws were present.

The flashlight is the most used tool in visual inspection. An improvement in the illumination uniformity of the beam has been suggested by installing a diffuser lens. We looked at inspection performance on a task of finding cracks from beneath fasteners where each inspector used a regular and a light shaping diffuser lens in a flashlight under different ambient light conditions. Under the conditions used, the enhanced lens did not hinder, but did not greatly enhance crack detection performance. It is possible that performances would be enhanced with tasks other than the one observed.

7. REFERENCES.

1. Shagam, R.N., Light Shaping Diffusers for Improved Visual Aircraft Inspection, DOT/FAA/AR95/32, October 1995.

2. Goranson, U.F. and Rogers, J.T., "Elements of Damage Tolerance Verification," 12th Symposium of International Commercial Aeronautical Fatigue, Toulouse, France, May 1983.

3. Bobo, S.N. and Puckett, C.H., Visual Inspection for Aircraft, Draft Advisory Circular AC 43-XX, FAA Aging Aircraft Program, Federal Aviation Administration, Atlantic City International Airport, New Jersey.

4. Latorella, K.A. and Drury, C.G., "A Framework for Human Reliability in Aircraft Inspection," 7th Federal Aviation Administration Meeting on Human Factors Issues in Aircraft Maintenance and Inspection, Atlanta, Georgia, August 1992.

5. Drury, C., "The Speed-Accuracy Tradeoff in Industry," *Ergonomics,* Vol 37, No. 4, pp. 747-763, 1994.

6. Drury, C.G. and Addison, J.L., "An Industrial Study of the Effects of Feedback and Fault Density in Inspection Performance," *Ergonomics,* Vol 16, pp. 159-169, 1973.

7. Wickens, C., Engineering Psychology and Human Performance, Scott, Foresman and Company, 1984.

8. Drury, C.G., "Exploring Search Strategies in Aircraft Inspection," in Brogan, D. (ed) Visual Search, Taylor and Francis Ltd., London, England, 1990.

9. Galwey, T. and Drury, C., "Task Complexity in Visual Inspection," *Human Factors,* Vol 28, pp. 595-606, 1986.

10. Noro, K., "A Descriptive Model of Visual Search," *Human Factors*, Vol 25, pp. 93-101, 1983.

11. Drury, C.G., "The Effect of Speed Working on Industrial Inspection Accuracy," *Applied Ergonomics*, Vol 4, pp. 2-7, 1973.

12. Newton, T. et al., "Personality and Performance on a Simple Visual Search Task," *Personality and Individual Differences*, Vol 13, pp. 381-382, 1992.

13. Megaw, E.D., "Factors Affecting Inspection Accuracy," *Applied Ergonomics*, Vol 10, pp. 27-32, 1979.

14. Splitz, G. and Drury, C., "Inspection of Sheet Materials: Test of Model Predictions," *Human Factors,* Vol 20, pp. 521-528, 1978.

15. Megaw, E. D. and Richardson, J., "Eye Movements and Industrial Inspection," *Applied Ergonomics,* Vol 10, pp. 145-154, 1979.

16. Kleiner, B., Drury, C., and Christopher, G., "Sensitivity of Human Tactile Inspection," *Human Factors,* Vol 29, No. 1, pp. 1-7, 1987.

17. Noro, K., "Analysis of Visual and Tactile Search in Industrial Inspection," *Ergonomics,* Vol 27, No. 7, pp. 733-743, 1984.

18. Konz, S. and Desai, S., "Tactile Inspection Performance With and Without Gloves," Human Factors 27th Annual Meeting: Turning the Tide of Technology, 1983.

19. Wilson, J. R., Corlett, E. N., The Evaluation of Human Work, A Practical Ergonomics Methodology, Taylor and Francis Ltd, 1990.

20. Drury, C.G. and Lock, M.W.B., "Ergonomics in Civil Aircraft Inspection," Contemporary Ergonomics—Proceedings of the Ergonomics Society's 1992 Annual Conference, Birmingham England, 7-10 April 1992.

21. Drury, C.G., "Errors in Aviation Maintenance: Taxonomy and Control," Proceedings of the Human Factors Society-35th Annual Meeting, 1991.

22. Rasmussen, J. and Vicente, K., "Coping with Human Errors Through System Design—Implications for Ecological Interface Design," *International Journal of Man Machine Studies,* Vol 31, No. 5, pp. 517-534, 1989.

23. Drury, C. and Lock, M., Reliability in Aircraft Inspection, UK and USA Perspectives, CAA Paper 94001, 1994.

24. Taylor, J., "Organizational Context for Aircraft Maintenance and Inspection," Proceedings of the Human Factors Society 34th Annual Meeting, 1990.

25. Proceedings of the 10th FAA/AAM Meeting on Human Factors in Aviation Maintenance and Inspection, Alexandria, Virginia, January 1996.

26. Endoh, S., Tomita, H., Asada, H., and Sotozaki, T., "Practical Evaluation of Crack Detection Capability for Visual Inspection in Japan," 17th Symposium International Committee on Aeronautical Fatigue, Stockholm, Sweden, 1993.

27. Drury, C.G., Improving Inspection Performance, Chapter 8.4 of Handbook of Industrial Engineering, G. Salvendy (ed), Wiley, NY, 1984.

28. Sheehan, J.J. and Drury, C.G., "The Analysis of Industrial Inspection," *Applied Ergonomics, B*, pp. 74-78, 1971.

29. Spencer, F.W. and Schurman, D.L., "Reliability Assessment at Airline Inspection Facilities, Vol III: Results of an Eddy-Current Inspection Reliability Experiment," DOT/FAA/CT-92/12,III, May 1995.

30. Nelson, L.S., "Sample Sizes for Two-Sample Tests on Means," *Journal of Quality Technology*, Vol 24, No. 2, pp. 103-108, 1992.

31. Drury, C.G. and Sinclair, M.A.. "Human and Machine Performance in an Inspection Task," *Human Factors,* Vol 25, No. 4, pp. 391-399, 1983.

32. Gramopadhye, A., Drury, C.G., and Prabhu, P.V., "Training for Visual Inspection of Aircraft Structures," Human Factors in Aviation Maintenance—Phase 3, Volume 1 Progress Report, DOT/FAA/AM-93/15, Springfield, VA, National Technical Information Service, August 1993.

33. Kleiner, B.M. and Drury, C.G., "Design and Evaluation of an Inspection Training Programme," *Applied Ergonomics*, Vol 24, No. 2, pp. 75-82, 1993.

34. Courtney, A.J., "A Search Task to Assess Visual Lobe Size," *Human Factors*, Vol 26, No. 3, pp. 289-298, 1984.

35. Patel, S., Prabhu, P., and Drury, C.G, "Design of Work Control Cards," Proceedings of the 7th FAA Meeting on Human Factors Issues in Aircraft Maintenance and Inspection, Atlanta, Georgia, August 1992.